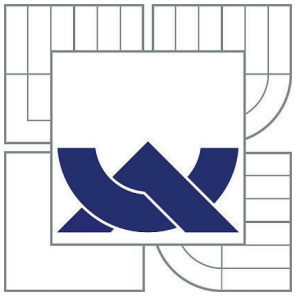
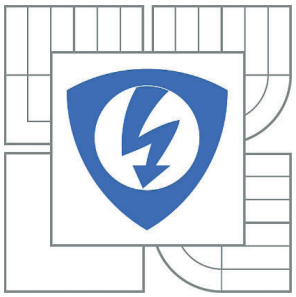


VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

ÚSTAV TELEKOMUNIKACÍ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

DEPARTMENT OF TELECOMMUNICATIONS

IDENTIFIKACE PAUZ V RUŠENÉM ŘEČOVÉM SIGNÁLU

IDENTIFICATION OF PAUSES IN NOISY SPEECH SIGNAL

DIPLOMOVÁ PRÁCE

MASTER'S THESIS

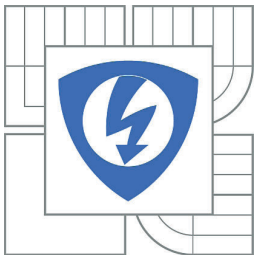
AUTOR PRÁCE — Bc. PETR KEPÁK

AUTHOR

VEDOUCÍ PRÁCE — prof. Ing. ZDENĚK SMÉKAL, CSc.

SUPERVISOR

BRNO 2011



VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

Fakulta elektrotechniky a komunikačních technologií

Ústav telekomunikací

Diplomová práce

magisterský navazující studijní obor
Telekomunikační a informační technika

Student:	Bc. Petr Kepák	***ID:***	21841
Ročník:	2	***Akademický rok:***	2010/2011

NÁZEV TÉMATU:

Identifikace pauz v rušeném řečovém signálu

POKYNY PRO VYPRACOVÁNÍ:

Základním problémem metod zvýraznění řeči je úplné oddělení přirozeného šumu, který vzniká při správně artikulaci znělých (sonorů) a neznělých souhlásek (konsonant) od šumu a rušení okolního prostředí. Cílem projektu je najít efektivní metodu, které by dokázala věrohodně identifikovat a oddělit nežádoucí šum a rušení od toho, který do řeči patří. Řešení problému souvisí s identifikací pauz bez řečové aktivity, v nichž je možné identifikovat vlastnosti šumu a rušení. Jakmile je správně šum určen, pak již je možné využít různých metod pro jeho odstranění.

DOPORUČENÁ LITERATURA:

[1] SMÉKAL, Z.: Číslicové zpracování signálu (MCSI). Elektronické učební texty pro magisterské studium, VUT Brno, 2009.
[2] PSUTKA, J., MULLER, L., MATOUŠEK, J., RADOVÁ, V.: Mluvíme s počítačem česky. Academia, Praha 2006. ISBN 80-2100-1309-1
[3] KRČMOVÁ, M.: Fonetika. Elektronické texty. MU Brno 2003.
http://is.muni.cz/do/1499/el/estud/ff/js07/fonetika/materialy/index.html

Termín zadání:	7.2.2011	***Termín odevzdání:***	26.5.2011

Vedoucí práce: prof. Ing. Zdeněk Smékal, CSc.

prof. Ing. Kamil Vrba, CSc.
Předseda oborové rady

UPOZORNĚNÍ:

Abstrakt

Základním problémem řeči je úplné oddělení přirozeného šumu, který vzniká při správné artikulaci znělých a neznělých souhlásek od šumu a rušení okolního prostředí.

Cílem diplomové práce je najít efektivní metodu, které by dokázala identifikovat pauzy bez řečové aktivity, v nichž je možné identifikovat vlastnosti šumu a rušení. Jakmile je správně šum určen, pak již je možné využít různých metod pro jeho odstranění.

V diplomové práci jsou popsány dvě metody identifikace pauz. Tyto metody jsou naprogramované v prostředí Matlab a testovány na devíti řečových nahrávkách. Analýza výsledků metod byla provedena pomocí ROC (Receiver Operating Characteristic) křivek.

V závěru jsou shrnuty výsledky analýzy vytvořených metod.

Klíčová slova

řeč, šum, pauza, střední hodnota, směrodatná odchylka, průchod signálu nulou, segmentace, ROC, FFT

Abstract

The basic problem of speech is a complete separation of the natural noise which arise from correct articulation of voiced and unvoiced consonants from noise and disturbance environment.

Objective of this master's thesis is to find an effective method that could identify the pauses without speech activity, which can identify the properties of noise and disturbance. Once the noise is correctly identified, it is already possible to use different methods for its removal.

The master's thesis describes two methods of pauses identification. These methods are programmed in Matlab and tested on nine speech recordings. Methods analysis of the results was performed using the ROC (Receiver Operating Characteristic) curves.

In the end are summarized results analysis of created methods.

Keywords

speech, noise, pause, mean, standard deviation, signal passes through zero, segmentation, ROC, FFT

KEPÁK, P. *Identifikace pauz v rušeném řečovém signálu.* Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, 2011. 65s. Vedoucí diplomové práce prof. Ing. Zdeněk Smékal, CSc.

Prohlášení o původnosti práce

Prohlašuji, že svou diplomovou práci na téma „Identifikace pauz v rušeném řečovém signálu“ jsem vypracoval samostatně pod vedením vedoucího diplomové práce s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny uvedeny v seznamu literatury na konci práce.

Jako autor uvedené diplomové práce dále prohlašuji, že v souvislosti s vytvořením této diplomové práce jsem neporušil autorská práva třetích osob, zejména jsem nezasáhl nedovoleným způsobem do cizích autorských práv osobnostních a jsem si plně vědom(-a) následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., včetně možných trestněprávních důsledků vyplývajících z ustanovení § 152 trestního zákona č. 140/1961 Sb.

V Brně dne ……... ……

podpis autora

Děkuji vedoucímu diplomové práce prof. Ing. Zdeňku Smékalovi, CSc. za velmi užitečnou metodickou pomoc a cenné rady při zpracování práce.

Dále bych rád poděkoval nejbližším za morální podporu nejen při vypracování této práce, ale i během celého studia.

V Brně dne ……... ……

podpis autora

Obsah

Seznam obrázků

Seznam tabulek

1. Úvod

V dnešní době plné různých komunikačních aplikací, sociálních sítí i jiných technických vymožeností stále zůstává základním a nejpřirozenějším prostředkem pro předávání informací mezi lidmi mluvená řeč.

Detekce řeči je významnou součástí mnoha aplikací pro zpracování řeči. Nachází využití v systémech pro zvýrazňování řeči k aktualizaci parametrů modelu pozadí řeči a také v řečových rozpoznávačích pro detekci začátku a konce promluvy a pro odstranění neřečových částí signálu.

Základním problémem řeči je úplné oddělení přirozeného šumu, který vzniká při správné artikulaci znělých a neznělých souhlásek od šumu a rušení okolního prostředí.

Cílem diplomové práce je najít efektivní metodu, které by dokázala identifikovat pauzy bez řečové aktivity, v nichž je možné identifikovat vlastnosti šumu a rušení. Jakmile je správně šum určen, pak již je možné využít různých metod pro jeho odstranění.

V diplomové práci jsou popsány dvě metody identifikace pauz. Tyto metody jsou naprogramované v prostředí Matlab a testovány na devíti řečových nahrávkách. Analýza výsledků metod byla provedena pomocí ROC (Receiver Operating Characteristic) křivek.

2. Definice řeči

Jazyk nám umožňuje předávat myšlenky prostřednictvím souboru znaků ať už grafických (latinka, řecká abeceda, azbuka, čínské ideogramy apod.), akustických (např. pomocí řeči), anebo jiných. Řeč je jedním z nejstarších a nejpřirozenějších prostředků komunikace mezi lidmi a je také jako prostředek komunikace nejčastěji užívána. [3]

Zvuková stránka lidského dorozumívání je předmětem zájmu po staletí. První popisy výslovnosti jsou známy z Indie, kde byly vytvořeny již v 7. století př. n. l.. Ze starých řeckých textů známe úvahy o hláskách, přízvuku, melodii řeči, Aristoteles v Rétorice (4. století př. n. l.) definuje hlásku, samohlásku, souhlásku i slabiku a podává třídění hlásek podle znělosti nebo místa artikulace.

V evropské lingvistice byly dlouhou dobu v popředí pozornosti hlásky chápané jako nejmenší jednotky mluvené řeči a zákonitosti jejich užívání, a to především v rámci slova. Nová vlna zájmu o zvuk řeči začíná v 19. století, ještě v polovině tohoto století se však poznatky o zvukové stránce jazyka mísívaly s fakty spojenými jen s psanými texty. Další vývoj lingvistiky dospívá rychle k osamostatnění pohledu, k vydělení fonetiky jako vědní disciplíny.

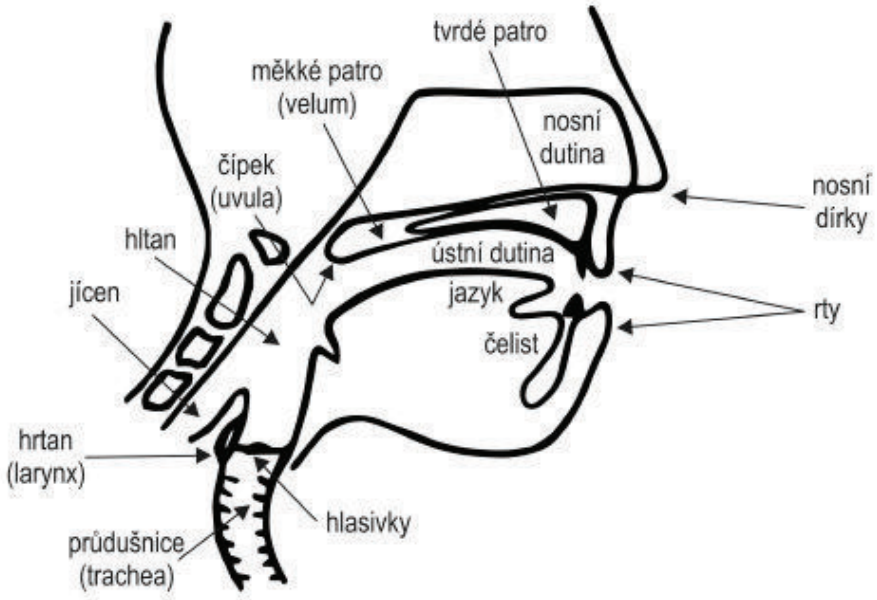
Mluvená řeč se přenáší komunikačním kanálem v podobě akustického signálu. Podstatou akustického (řečového) signálu je vlnění elastického prostředí v množině slyšitelných frekvencí.

Pod pojmem komunikační kanál si představme prostředí, kterým se šíří akustický signál od zvukového ústrojí mluvčího ke sluchovému ústrojí posluchače. Kromě akustické složky (amplitudově-frekvenční časové spektrum) řečový signál obsahuje lingvistickou složku (fonetická, morfologická, syntaktická, sémantická či pragmatická struktura) vyjadřující význam promluvy.

2.1 Proces vytváření řeči člověkem

V lidském těle je několik orgánů, které se zabývají vytvářením řeči, souhrnně tyto orgány nazýváme řečové (artikulační) orgány, nebo také jednoduše mluvidla (artikulátory). Seskupení těchto orgánů v těle tvoří hlasový trakt. Hlasový trakt lze rozdělit na tři základní části [1]:

- dechové ústrojí
- hlasové ústrojí
- artikulační ústrojí



[14]

Obr. 2.1: Hlasový trakt

2.1.2 Dechové ústrojí

Dechové ústrojí představuje zdroj energie pro řeč. Je umístěno v hrudním koši a tvořeno přívodní dýchací cestou, plícemi a s nimi funkčně spjatými dýchacími svaly (bránicí). Při nádechu dochází k pohybu vzduchu, který tak poskytuje zdroj energie pro řeč. Při výdechu potom v plicích vzniká výdechový proud vzduchu, který je v zásadě základním materiálem pro tvorbu řeči.

Výdechový proud je z plic odváděn průdušnicí, a pak prochází hrtanem a nadhrtanovými dutinami, kde se modifikuje a jako řečový signál je vyzařován rty do okolního prostoru. Síla výdechového proudu ovlivňuje způsob fungování hlasového ústrojí, a tím má vliv na sílu hlasu a částečně i na jeho výšku.

2.1.3 Hlasové ústrojí

Pojmem hlasové ústrojí se často označuje celý systém pro vytváření řeči. Zde budeme pod tímto pojmem rozumět pouze tu část, kde bude docházet k samotnému vzniku hlasu. Hlasové ústrojí je uloženo v hrtanu, který je s plícemi spojen průdušnicí. Z hlediska tvorby řeči nejdůležitější část hlasového ústrojí tvoří hlasivky. Jsou to dvě ostré slizniční řasy, které vedou napříč hrtanem v místě jeho nejužšího průchodu.

Při vytváření hlasu (fonaci) se hlasivky nacházejí v tzv. hlasovém (fonačním) postavení. Výdechový proud vzduchu postupuje bez odporu z plic průdušnicí až k hrtanu. Zde se mu do cesty postaví překážka vytvořená hmotou hlasivek, které cestu vzduchu úplně uzavřou. Stažené hlasivky se pod tlakem vzduchu stávají pružnými a začínají kmitat. V důsledku kmitání hlasivek vzniká základní (hlasivkový) tón, který představuje nosný zvuk řeči.

Fonační postavení hlasivek má za následek vznik hlasivkového tónu a používá se proto při vytváření znělých zvuků řeči (samohlásky a znělé souhlásky). Neznělé zvuky jsou naopak tvořeny při klidovém postavení hlasivek, neobsahují tedy základní hlasivkový tón a vznikají tedy až modifikací výdechového proudu vzduchu v nadhrtanových dutinách.

2.1.4 Artikulační ústrojí

Artikulační ústrojí je posledním ústrojím, které se podílí na tvorbě řeči. Jeho význam spočívá v tom, že umožňuje vytvářet velké množství různých zvuků, které charakterizují mluvený jazyk. Skládá se jednak z nadhrtanových dutin a jednak z artikulačních orgánů, které jsou v těchto dutinách uloženy nebo je obklopují. Mezi nadhrtanové dutiny řadíme dutinu hrdelní, ústní a nosní.

Hranici mezi těmito dutinami tvoří čípek, špička měkkého patra, které zamezuje nebo umožňuje přístup vzduchu z dutiny hrdelní do dutiny nosní. Zatímco se nadhrtanové dutiny účastní procesu tvorby řeči pasivně (nepohybují se), artikulační orgány (artikulátory) se účastní tvorby řeči většinou aktivně – tvoří pohyblivé součásti artikulačního ústrojí a svým pohybem mění velikost nadhrtanových dutin.

Z hlediska vytváření řeči mezi nejvýznamnější artikulátory patří jazyk, rty a měkké patro, neboť se podílejí na vytváření největšího počtu různých zvuků. Dalšími artikulátory potom jsou zuby, tvrdé patro nebo čelisti. Artikulátorem je také hrtan, který se může pohybovat a tím měnit délku celého hlasového traktu.

2.2 Vlastnosti a skladba řeči

2.2.1 Základní tón lidského hlasu

Frekvence kmitání hlasivek se označuje f_0 a nazývá se frekvence základního tónu hlasu. Tato frekvence nabývá hodnot asi od 60−400 Hz. U mužů se f_0 pohybuje asi mezi 80−160 Hz, u žen je to 150−300 Hz a u dětí asi 200−400 Hz.

Základní tón hlasu je přítomen při tvoření všech znělých zvuků, tj. samohlásek a znělých souhlásek.

2.2.2 Slovo

Slovo funguje v promluvě jako jednotka významová. V jazycích, které mají člen, vstupuje slovo do textu většinou s tímto členem, často se obě složky i vážou natolik, že pojmenovávací jednotka (běžný význam termínu ´slovo´) se se zvukovou jednotkou nekryje. Nicméně v mateřském jazyce poznáváme slovo na základě komplexu vlastností významových, formálních i zvukových. Jednoznačné zvukové kritérium hranice slova neexistuje, a to ani v češtině, která má obvykle přízvuk na první slabice - mnoho slov jej totiž v textu nemá.

Hranice slova se mohou v jednotlivých jazycích signalizovat i dík jistému omezení ve využití hlásek nebo jejich skupin na počátku nebo na konci slova nebo dík omezení v přizpůsobování hlásek (jevy koartikulace, např. splývání hlásek patřících různým slabikám, se mnohdy v ortoepické výslovnosti realizují jen v rámci slova). V tomto smyslu však jde jen o poměrně slabé zvukové signály. Slovo je zvukově vydělováno také na základě jeho potenciální schopnosti být samostatným taktem. Jeho slabičný sklad je v jazyce ustálený.

Na slova jako na fonetické útvary lze pohlížet jako na spojení konečného počtu fonémů či slabik.

Výhodou užití slova jako segmentační jednotky je eliminace nutnosti zacházet s komplikovanými algoritmy pro segmentaci a identifikaci nižších jednotek. Tato výhoda se projeví zejména v případě, že slova jsou vyslovována izolovaně, tj. jsou obklopena dostatečně dlouhými pauzami (alespoň 0,2 až 0,3 s.). V tomto případě lze nalézt hranice slov relativně snadno s ohledem na množství šumu ve vzorku.

Slovanské jazyky používají kolem 2 500−3 500 slabik a 45 000−50 000 slov.

2.2.3 Pauza

Pauza je přerušení řečového proudu. Vzniká z fyziologických příčin tam, kde je třeba doplnit dech. Pro sdělování má význam komunikativní pauza oddělující větší řečové celky. Z hlediska trvání jde buď o pauzu absolutní (má neomezenou délku), nebo relativní (krátkodobé přerušení řeči).

Pro fungování a porozumění řeči má význam to, že pauza je někdy jen potenciální, tj. může (ale nemusí) být v některé pozici realizována, její realizace je jen fakultativní (nezávazná).

Potenciální pauza je realizována podle konkrétní situace, důležitá je jen pozice, kde může být; tak může být např. přerušen řečový proud mezi slovy, ne však mezi jednotlivými hláskami; to už by se ztrácela souvislost řeči. Potenciální pauza pak slouží při komunikaci jako prvek ukazující hranici řečových celků.

2.2.4 Slabika

Slabiky jsou fonetické útvary, které obsahují samohláskové jádro a volitelné počáteční a koncové souhlásky nebo jejich skupinu. Slabika tak obsahuje jak přechody souhláska-samohláska, tak i přechody samohláska - souhláska, včetně koartikulací a jiných fonologických efektů uvnitř jejich hranic. Slabiku tvoří její jádro, nukleus, ohraničené slabičnými svahy (iniciálou tzv. „onsetem“ před a kodou za jádrem). Slabiky spojujeme na fonetické úrovni v takty sdružené jedním přízvukem.

Slabika (sylaba) je nejjednodušší a nejtěsnější možnou artikulační jednotou funkčních prvků řeči, která vyhovuje dorozumívání. Členění slov na slabiky odpovídá i přirozenému jazykovému citu uživatelů, který se kultivuje už v raném dětství (odříkávání rozpočítadel, říkadel apod.).

Hláska je laicky pokládaná za základ zvukové řeči - jednotkou hypotetickou, jež ve skutečné řeči (pokud náhodou sama není slabikou - např. spojky a, i v češtině) jako samostatná výslovnostní jednotka neexistuje. Pokud "hláskujeme", vlastně vytrháváme na základě empirie zvuky a u souhlásek je navíc nutně doprovodíme neurčitou samohláskou.

Pocit slabičného skladu mluvy, který všichni uživatelé mají, vychází zjevně z artikulačních a akustických skutečností, jež slabiku charakterizují. Fonetickou podstatu slabiky se však dosud nepodařilo jednoznačně popsat: vždy zůstává jistý počet hláskových kombinací, jež jsou uživateli hodnoceny jako slabiky, ale teoretickým vymezením neodpovídají. Žádný z principů vymezení popsaných ve starší literatuře nedostačuje sám pro poznání podstaty slabiky, neboť při artikulaci i vnímání se uplatňují společně a v jednotlivých typech slabik vždy některý vystupuje do popředí.

2.2.5 Hláska

Za nejmenší jednotku řeči, která může rozlišovat jednotlivá slova, lze považovat hlásku - foném. Fonémy lze od sebe rozlišit například podle místa tvoření, podle artikulujícího orgánu, nebo podle sluchového dojmu.

Počet fonémů v existujících světových jazycích se pohybuje od 12 do 60. V českém jazyce je jich 36, v anglickém 42, v ruském 40 apod. Fonémy se spojují do posloupnosti spojených celků, v nichž lze nalézt další stavební jednotku – slabiku

Z akustického hlediska je nejvhodnější rozdělit hlásky na: [2]

Vokály (vocoidy, **samohlásky**) - hlásky s volným vyzněním hlasu doplněným rezonancemi. Artikulačně jde o hlásky založené na apertuře. Do této skupiny patří běžně jak monoftongy, tak složitější polyftongy.

Při artikulaci samohlásek je průchod vzduchu hlasovým traktem poměrně volný. Samohlásky zachovávají v časovém průběhu kvaziperiodicitu, která se v kmitočtové oblasti projeví soustředěním většiny energie do několika kmitočtových pásem, tzv. formantových pásem. Formanty jsou kmitočty odpovídající rezonančním kmitočtům jednotlivých dutin hlasového traktu člověka. Český jazyk má pět samohlásek (a, e, i, o, u).

Konsonanty (contoidy, **souhlásky**) - hlásky, jejichž zvukový obraz je založen na šumu (samostatném nebo doplněném i složkou tónovou). Artikulačně jde o hlásky, v jejichž tvoření je uplatňuje striktura.

Souhlásky vznikají jako šelesty při proudění vzduchu skrz zúžená místa (například souhláska „s“ je soubor velmi vysokých tónů, vznikajících při proudění vzduchu mezi zuby) nebo tím, že rty, zuby nebo jazyk náhle otevírají cestu pro vzduch proudící z plic, čímž vznikají jen krátce trvající nepravidelné zvuky.

Souhlásky mají šumový charakter a až na výjimky hlasivky nekmitají. Souhlásek je podstatně více než samohlásek a proto jsou ještě dále členěny na párové znělé, párové neznělé, nepárové, závěrové, úžinové a polozávěrové.

Lze je dělit dle typu překážky:

- překážka může být úplná, pak vznikají tzv. **závěrové** souhlásky (p, t, ť, k, b, d, ď, g, m, n, ň)
- jiným typem překážky je zúžení, kdy vzniká charakteristický třecí šum. Souhlásky takto tvořené se nazývají **úžinové** (f, v, s, š, z, ž, j, ch, h, l, r, ř)

- u malé skupiny hlásek se vyskytují postupně oba typy překážek, tyto souhlásky se nazývají **polozávěrové** (c, č)

Také lze zavést dělení souhlásek dle párovosti:

- většinu šumových souhlásek je možné seřadit do dvojic, v nichž se obě souhlásky v podstatě shodují co do způsobu artikulace, liší se však znělostí - takové souhlásky jsou **párové** (p-b, t-d, ť-ď, k-g, s-z, š-ž, f-v, ch-h, c-dz, č-dž)
- souhlásky **nepárové**, které jsou vždy znělé (m, n, ň, l, j, r, ř)

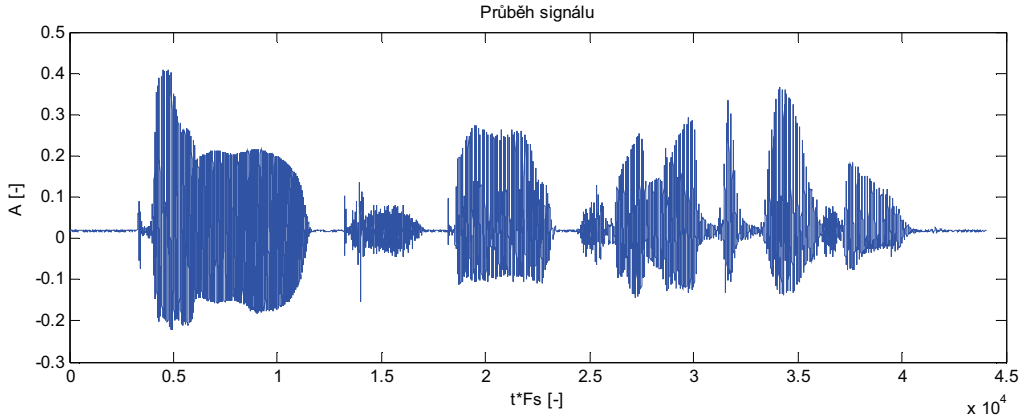
Početná skupina souhlásek není z akustického hlediska stejnorodá, podle vztahu mezi hlasovou a šumovou složkou je lze dále dělit na:

- **Hlásky klouzavé (glide)**, jež nemají plně rozvinutou tónovou složku, ale chybí jim i výraznější složka šumová. Tyto hlásky mají poměrně nejblíže k vokálům, v některých jazycích se dokonce v jednotlivých slovech se skutečným vokálem střídají. Artikulačně jde o aproximanty, v lingvistice se o nich mluví také jako o semivokálech, polovokálech
- **Sonory** - hlásky s poměrně větším podílem tónové složky, ale obsahující i složku šumovou; na jejich akustickém obrazu se vedle hlasu podílí i rezonance: dělí se na likvidy (souhlásky plynné, [l], [r], u nichž tónovou složku vytváří rezonance v ústech), a nazály, kde je doplňující rezonance nosní (např. [m], [n]). Blízkost těchto souhlásek, zvláště likvid, k vokálům ukazuje jejich schopnost fungovat v některých jazycích jako jádro slabiky (tak i v češtině - např. jednoslabičná slova vlk, prst). Označení "sonora", "likvida" jsou typickým termínem založeným na sluchovém dojmu, užívají se však i tam, kde se preferuje pojmenovaní na základě artikulace nebo akustiky.
- **Vlastní konsonanty** (konsonanty šumové, obstruenty) mají zřetelnou složku neperiodických kmitů (šumů). Částečně tónovou složku mají z této skupiny konsonanty znělé - tvoří ji znění základního hlasu), konsonanty neznělé jsou čistými šumy.

3. Základní charakteristiky řečového signálu

Metody krátkodobé analýzy vycházejí z podstaty kvazistacionarity řeči. Hlasové ústrojí je totiž schopné měnit své parametry vždy až po určitém čase, v němž se řeč jeví jako stacionární a ergodický signál. Tato délka je mezi 10 až 35ms. Proto je signál pro potřeby analýzy a zpracování rozdělen do stejných časových segmentů takto definované délky.

V technikách zvýrazňování řeči se často používají funkce, které popisují vlastnosti řečového signálu. Mezi základní charakteristiky patří krátkodobá energie, krátkodobá intenzita a autokorelační funkce. Hodnoty těchto charakteristik mohou být vypočítány v každém časovém segmentu signálu a jejich dlouhodobé sledování je základem pro kategorizaci řečového signálu, výpočty odhadu šumu nebo predikce následujícího chování signálu. Neumožňují však zpětnou rekonstrukci signálu. [11]

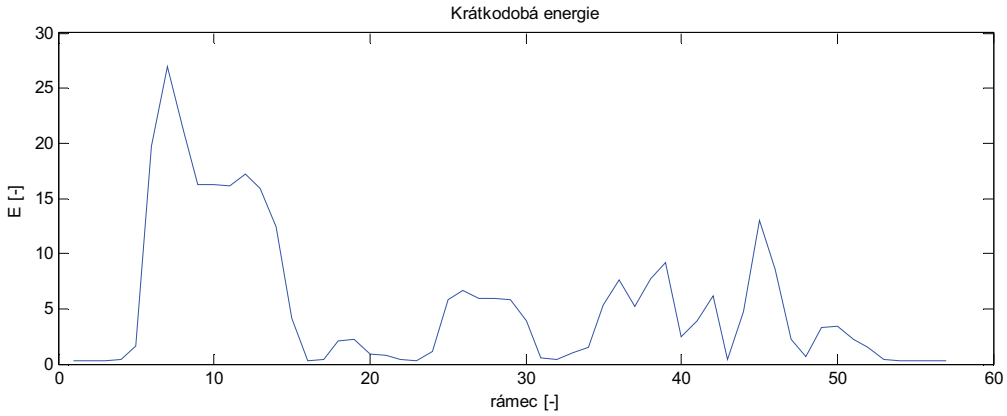


Obr. 3.1: Průběh ukázkového signálu

Krátkodobá energie je jedna ze základních funkcí, kterou lze pro N vzorků definovat vztahem

$$E = \sum_{n=0}^{N-1} x[n]^2 \tag{3.1}$$

kde $x[n]$ je signál v čase *n*.

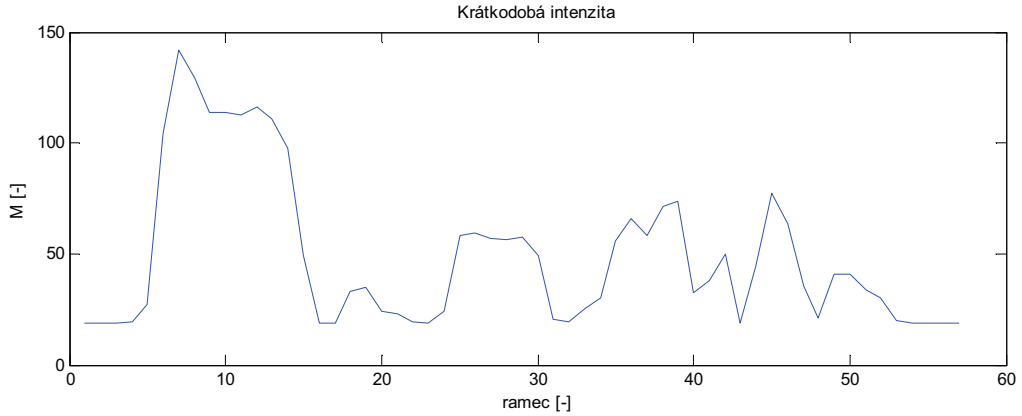


Obr. 3.2: Krátkodobá energie

Krátkodobá intenzita se často používá jako alternativa k funkci krátkodobé energie, protože ta má díky druhé mocnině velkou číselnou dynamiku. Krátkodobá intenzita je definována rovnicí

$$M = \sum_{n=0}^{N-1} \left| x[n] \right| \tag{3.2}$$

Obě funkce jsou využívány v energetických detektorech řečové aktivity a pro určení segmentů znělých a neznělých částí promluvy.



Obr. 3.3: Krátkodobá intenzita

Autokorelační funkce určuje míru vzájemné podobnosti jednotlivých vzorků signálu. Může se vypočítat podle vzorce

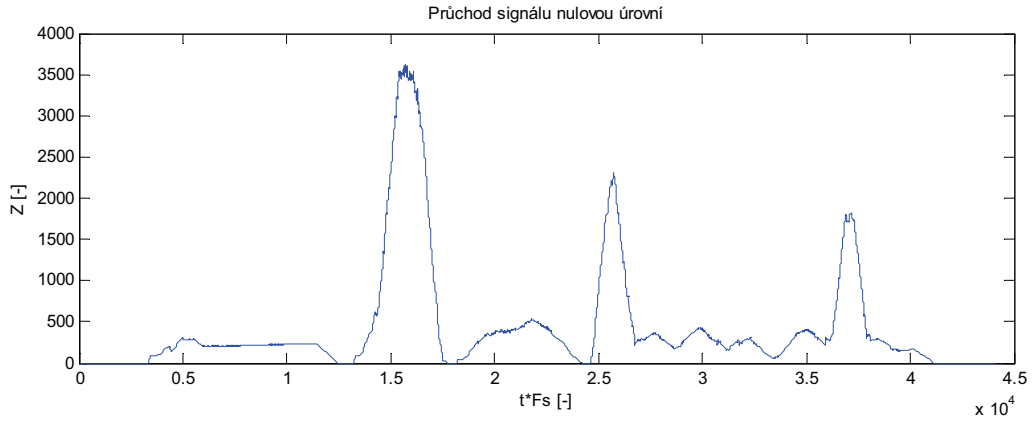
$$R[m] = \frac{1}{N} \sum_{n=m}^{N-1} x[n] \cdot x[n-m] \tag{3.3}$$

Pokud je zpracovaný signál periodický s periodou *T* nabývá tato funkce lokálních maxim právě při *m* = 0,*T*,2*T*,…. Tato charakteristika umí rozlišit deterministický nebo náhodný signál. Používá se pro určení korelovanosti šumu nebo při některých výpočtech spektrální výkonové hustoty.

Průchod signálu nulovou úrovní se využívá k určení znělých a neznělých hlásek. Při vytváření neznělých hlásek se nepoužívají hlasivky, proto se více podobají šumu a mají tedy vyšší Z. Počet průchodů nulou je ale značně náchylný na šum, je proto někdy problém rozlišit neznělou hlásku od šumu.

$$Z = \frac{1}{2}\sum_{n=1}^{ramec} \left| signx[n] - signx[n-1] \right| \tag{3.4}$$

kde $\begin{Bmatrix} sign\ x[n] = 1 & pro\ x[n] \succ 0 \\ sign\ x[n] = -1 & pro\ x[n] \prec 0 \\ sign\ x[n] = 0 & pro\ x[n] = 0 \end{Bmatrix}$



Obr. 3.4: Průchod signálu nulovou úrovní

4. Analýza šumů

Zašuměný řečový signál se může definovat jako součet řeči a šumového signálu. Samotné šumy a rušení se rozdělují do dvou základních skupin, na šumy aditivní a konvoluční. [12]

Konvoluční šum souvisí s užitečným signálem. Nejčastěji vzniká už při zachycení zvuku odrazem od okolních překážek, dále v průběhu přenosové cesty jako zkreslení způsobené odrazy nepřizpůsobeným vedením nebo časovým zpožděním, anebo na nelineárních aktivních prvcích. Tento druh zkreslení, pokud už vznikne, se odstraňuje jinými metodami.

4.1 Aditivní šum

Aditivní šum je skupina šumů, které se k užitečnému signálu přičtou v akustické rovině při snímání akustickými snímači, nebo v elektrické rovině přidanými šumy aktivních prvků nebo přeslechy z jiných vedení bez souvislosti s přenášeným řečovým signálem. Aditivní šum není korelovaný s řečovým signálem. Aditivní šumy se mohou rozdělit na několik skupin podle barvy, korelace, obsažené šířky ve spektru, nebo podle průběhu a chování v čase.

4.1.1 Bílý a barevný šum

Bílý gaussovský šum je příkladem širokopásmového rušení. Je popsán rozptylem σ_n^2 a střední hodnotou μ_n. Důležitou charakteristikou je spektrální výkonová hustota, která je pro $\mu_n = 0$ konstantní a platí $P[n] = \sigma_n^2$. Bílý šum je nekorelovaný, stacionární a nenese žádnou informaci. Aproximací reálného šumového pozadí je barevný šum, který nemá konstantní amplitudu spektrálních čar.

4.1.2 Korelovaný a nekorelovaný aditivní šum

Korelace šumu určuje vzájemnou podobnost jednotlivých vzorků šumu v čase autokorelační funkcí.

$$R[m] = \frac{1}{N} \sum_{n=m}^{N-1} x[n] \cdot x[n-m] \tag{4.1}$$

Pro korelovaný šum platí $R[m] \neq 0$. Barevný šum je tedy korelovaný. Nekorelovaný šum má $\lim_{N \to \infty} R[m] = 0$ pro m>0 a jeho typickým případem je bílý šum. Korelovaný aditivní šum není korelovaný s užitečným řečovým signálem.

4.1.3 Širokopásmový šum

Aditivní šum se může dělit také podle obsahu spektrálních čar. Jestliže je energie obsažena ve všech uvažovaných spektrálních čarách, jedná se šum širokopásmový. Typickým příkladem je bílý šum. Jestliže je obsažena jen v některých frekvenčních pásmech, jedná se o šum úzkopásmový. Příkladem může být síťový brum ve zvukovém kanálu.

4.1.4 Stacionární šum

Stacionární šum má v definovaném čase konstantní hodnoty vyhlazeného spektra. V metodách krátkodobé analýzy je zvolen časový segment, ve kterém se předpokládá, že šum i řeč je v tomto segmentu stacionární. Jestliže má následující segment odlišnou spektrální charakteristiku, jedná se o kvazistacionární signál, vzhledem k délce časového segmentu. Nestacionární řečový signál je v tomto pojetí kvazistacionární. Bílý a barevný šum má hodnoty vyhlazeného spektra pro každý segment konstantní, je tedy stacionární. Šum nestacionární (i řeč) má vyhlazené spektra v časových segmentech odlišné.

4.2 Reálné aditivní šumy v řečovém signálu

Reálné aditivní šumy mohou mít různé vlastnosti. Většinu lze rozdělit na složku stacionární a složku nestacionární. Stacionární složku šumu aproximuje většinou barevný šum, nebo konkrétní korelované frekvence síťového rušení nebo hluku ložisek točivých strojů. Nestacionární složku představuje šum způsobený nepravidelnými rázy, větrem, hudbou nebo jinými mluvčími. Nestacionární nekorelovanou složku aditivního šumu tvoří impulsní rušení jako jsou elektrické nebo akustické poruchy.

Řečový signál je kvazistacionární, kdy v časových segmentech do délky 30ms se jeví jako stacionární. V metodách zvýrazňování řeči se předpokládá, že šum má charakter stacionární nebo kvazistacionární, a že rychlost změny šumu je podstatně nižší než rychlost změny řeči. Při rychlejší změně šumu, než je schopnost aktualizace odhadu šumové složky (výpočet vyhlazeného odhadu), vznikají vysoké residuální šumy ve zvýrazněné řeči. Nestacionární složka šumu se eliminuje jinými technikami. [11]

5. Extrakce příznaků řeči

5.1 Převod analogového signálu na digitální signál

5.1.1 Vzorkování

Vzorkovací frekvence (*fs*) udává, kolikrát za 1 sekundu provádíme měření. CD je vzorkováno na 44,1 kHz, řeč mezi 16 a 32 kHz.

Vzorkovací teorém - *fs* musí být vyšší než dvojnásobek nejvyšší frekvenční složky signálu. Na jednu periodu nejvyšší harmonické musejí připadnout alespoň dva vzorky.

5.1.2 Kvantizace

Velikost intervalu amplitudy - čím jsou intervaly menší, tím je měření přesnější. Každý vzorek je kvantizován pomocí binárních jednotek bitu. Čísla 0 a 1 lze uložit do jednoho bitu, čísla 0-3 potřebují dva bity (4 hodnoty), čísla 0-7 tři bity.

Bitové rozlišení udává, na kolik intervalu celý rozsah amplitudy vzorkujeme. 8bitové rozlišení znamená 256 možných intervalu, 16bitové rozlišení znamená 65536 intervalu.

5.2 Preemfáze

Před vlastním zpracováním řečového signálu se často využívá tzv. preemfáze. Preemfáze znamená zdůrazňování amplitud spektrálních složek řečového signálu s jejich vzrůstající frekvencí. Důvod tohoto procesu vyplývá z opačného chování řečového ústrojí, tj. poklesu amplitud spektrálních složek řečového signálu na vyšších frekvencích.

Preemfáze tento pokles v jisté míře kompenzuje, takže dojde k relativnímu vyrovnání energetického spektra celého přenášeného pásma. Pro číslicové zpracování řeči může být preemfáze realizována dvěma způsoby:

- analogový filtr, jehož frekvenční charakteristika má strmost +20 dB/dek od lomové frekvence 100 Hz,
- číslicový filtr horní propusti prvního řádu typu FIR, který provádí předzpracování signálu podle vztahu:

$$y[n] = x[n] - \mathrm{a}x[n\text{-}1] \tag{5.1}$$

kde *x(n)* vstupní vzorek v čase
y(n) výstup v čase
konstanta *a* se volí v intervalu (0,9 ; 1)

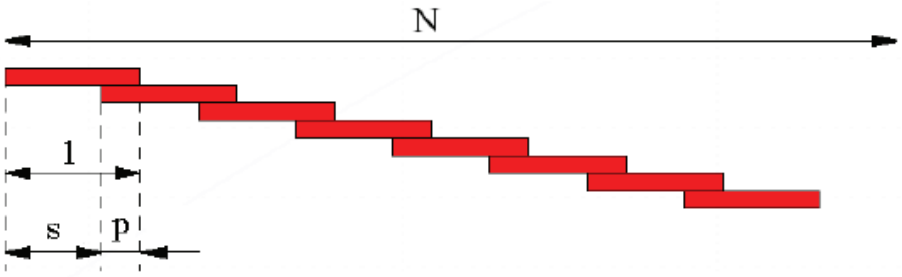
5.3 Rámce

Řečový signál je nutné před dalším zpracováním rozdělit na rámce. Důvod je prostý. Metody pro odhad parametrů dobře pracují se stacionárním signálem, což celý signál určitě není. Pokud chceme zpracovávat stacionární úsek, musíme uvažovat setrvačnost hlasového ústrojí. Signál tedy většinou dělíme na 20−25 ms dlouhé úseky. Délka rámce ve vzorcích při 16 kHz je 320−400 vzorků.

Pro lepší zachování kontextu je vhodné, aby se jeden úsek řeči stal součástí několika rámců, tedy aby se sousední rámce z určité části překrývaly. Samozřejmě s rostoucí hodnotou překrytí se zvyšují i nároky na paměť a také si budou sousední rámce více podobné. Nicméně s malým překrytím se hodnoty parametrů mezi sousedními rámci můžou hodně měnit. Proto je vhodné udělat kompromis, kterým je z pravidla překrytí 10-15 ms.

Samotné rámcování signálu se provádí pomocí vykrojení tzv. okna. Tvar okna může být různý, nejčastěji se ovšem využívá pravoúhlé nebo Hammingovo okno. [13]

V této práci se dělení provádí s rámci délky 32 ms a 64ms s překrytím rámců o 1/4.



Obr. 5.1: Překrývání rámců

kde *l* délka jednoho rámce

N celkový počet navzorkovaných hodnot

p velikost překrytí rámce

s posuv

5.4 Okenní funkce

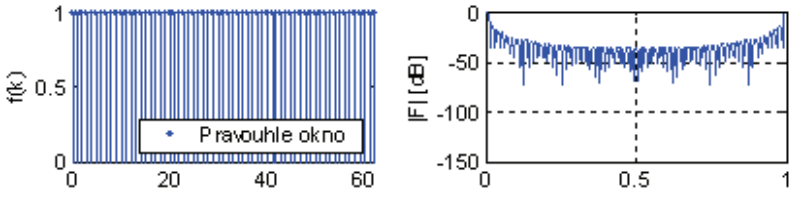
Existuje mnoho typů okenních funkcí, které mají různé filtrační vlastnosti. Především se jedná o šířku hlavního laloku a potlačení takzvaných postranních laloků. Čím širší je hlavní lalok, tím je vyšší přelévání energie spektra a tedy nižší „ostrost“ spektra. A čím je hlavní lalok okenní funkce užší, tím je více postranních laloků a tedy více falešných frekvencí ve spektru.

5.4.1 Pravoúhlé okno

Základní typ okna, jednoduše vyřízneme část signálu

Pro počítač to znamená, že signál náhle začíná a náhle končí - silné vysokofrekvenční složky. Proto se používá různých algoritmu pro stanovení jiných tvaru oken („vyhlazení rohu“). Na okno „přiložíme“ křivku, která stoupá od 0 k 1 a zase klesá zpět na 0, a vlnu jí vynásobíme

Pravoúhlé okno má nejmenší potlačení nepropustného pásma a velmi úzké propustné pásmo. Z toho vyplývá, že spektrální rozlišení je při váhování pravoúhlým oknem velmi dobré.



[5]

Obr. 5.2: Pravoúhlé okno: časový průběh a normované amplitudové spektrum

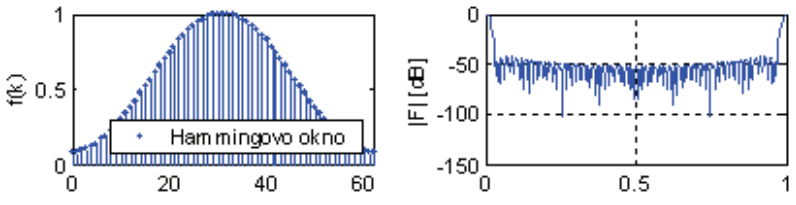
5.4.2 Gaussovo okno

Je nejvýhodnější, protože dokáže blokovat „postranní sukně“ (laloky)

5.4.3 Hammingovo okno

Hammingovo okno, které utlumí signál na okrajích rámce a zabrání tak rušivým přechodovým jevům. Hammingovo okno má téměř konstantní potlačení v celém nepropustném pásmu. Používá se v krátkodobé analýze při použití techniky sčítání přesahů s 50% přesahem, kde vykazuje nejnižší zvlnění signálu po rekonstrukci. Hammingovo okno délky N je definováno takto:

$$w(n) = 0{,}54 - 0{,}46\cos\left(\frac{2\pi n}{N}\right) \qquad 0 \le n \le N-1 \qquad (5.2)$$



[5]

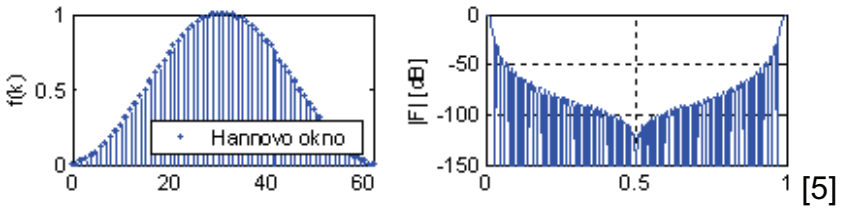
Obr. 5.3: Hammingovo okno: časový průběh a normované amplitudové spektrum

5.4.4 Bartlettovo okno (trojúhelníkové)

Bartlettovo okno se používá zřídka, protože jeho kmitočtové vlastnosti nejsou nijak oslnivé.

5.4.5 Hannovo okno

Potlačení Hannova okna roste. Je to však za cenu dvojnásobné šířky nepropustného pásma (oproti pravoúhlému) a tudíž i horšího spektrálního rozlišení.



[5]

Obr. 5.4: Hannovo okno: časový průběh a normované amplitudové spektrum

5.4.6 Hanningovo okno

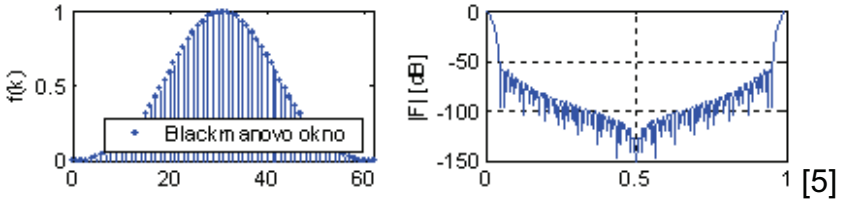
Hanningovo okno je často používaná okenní funkce s dobrým kompromisem mezi ostrostí spektra a potlačení falešných frekvencí. Má úzký hlavní lalok a malé postranní laloky. Používá se v úlohách krátkodobé analýzy.

$$w(n) = \frac{1}{2}\left(1 - \cos\frac{2\pi n}{N}\right) \qquad 0 \le n \le N-1 \qquad (5.3)$$

5.4.7 Blackmanovo okno

Blackmanovo okno má v některých ohledech lepší a v některých horší kmitočtové vlastnosti než nejužívanější Hammingovo okno. Blackmanovo okno má největší potlačení nepropustného pásma, avšak nehorší spektrální rozlišení.

$$w(n) = 0{,}42 + 0{,}5\cos\left(\frac{2\pi n}{N}\right) + 0{,}08\cos\left(\frac{4\pi n}{N}\right) \qquad 0 \le n \le N-1 \qquad (5.4)$$



[5]

Obr. 5.5: Blackmanovo okno: časový průběh a normované amplitudové spektrum

6. Databáze řečových nahrávek

Za účelem testování algoritmů zpracování řeči jsou vytvářeny databáze multilingválních nahrávek promluv. Nahrávky jsou pořizovány pro mužské, ženské, ale i dětské mluvčí jako čisté nebo zašuměné. Různé druhy šumů jsou do nahrávek přidávány buď aditivně, nebo jsou promluvy zaznamenány přímo v prostředí, kde je šum přítomen. Z hlediska ověření účinnosti nových algoritmů je poslední možnost nejvhodnější, neboť takto získaný audio signál představuje reálné prostředí. K nahrávkám jsou přidávány ortografické informace, fonetická transkripce, nebo indexy poloh jednotlivých úseků řeči získaných manuální segmentací v textové podobě. Zpravidla se jedná o hranice fonémů, slabik, slov a pauz v milisekundách. Vzájemným porovnáním originální databáze a výstupu algoritmů na ní aplikovaných je získána jejich účinnost.

Databázi tvoří vyprávění devíti italských mluvčí. Všechny nahrávky se vztahují ke shodné scéně kresleného filmu, jež mluvčí viděli, a interpretují přátelům. Za tímto účelem byli nahráváni na digitální videokameru pokaždé v různém prostředí s různými druhy šumů na rozdílných úrovních. Audio nahrávky se vzorkovací frekvencí 32kHz a 16kHz, kvantováním 16 bitů a minimální průměrnou hodnotou SNR 3 dB byly extrahovány pro testovací účely přímo z videa. Databáze vznikla především za účelem detekce pauz a nelexikálních prvků, a proto byly tyto úseky řeči manuálně segmentovány (s odbornou korekcí) pomocí programu Speechstation27. [10]

V diplomové práci je použita databáze devíti řečových nahrávek, která tvoří základ pro testování VAD algoritmů.

Tab. 6.1: Databáze řečových nahrávek

Název nahrávky	Délka nahrávky	F_{vz} [kHz]	stereo / mono
Alessio.wav	3:05	32	stereo
Antonio.wav	2:37	32	stereo
Carmine.wav	1:39	32	stereo
Franco.wav	5:39	16	mono
Gerry.wav	7:05	16	stereo
Marco_magliuolo.wav	2:18	32	stereo
Marco_riccio.wav	1:37	32	stereo
Paolo.wav	1:50	32	stereo
Tonyf.wav	5:55	16	mono

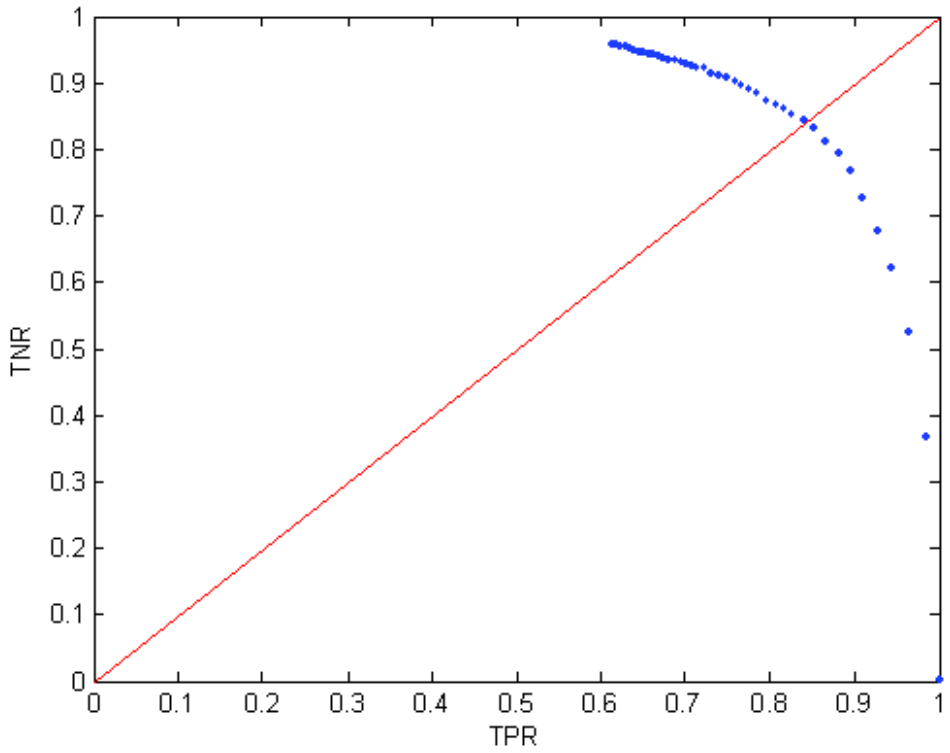
7. ROC analýza

ROC (Receiver Operating Characteristic) analýza poskytuje nástroje pro výběr optimální metody nebo pro výběr optimálního řešení. Je to statistický postup pro vyhodnocení signálů správné a falešné pozitivity a správné a falešné negativity.

Má původ v hodnocení kvality signálu britských radarů za druhé světové války (bylo nutné rozpoznat správně pozitivní signál od falešně pozitivních signálů). Následně byla analýza použita v psychologii pro detekci vnímání podnětů. ROC analýza se od té doby používá v medicíně, radiologii a dalších oblastech po mnoho desetiletí.

7.1 ROC křivka

ROC křivka je někdy také označována jako hraniční křivka. Je to křivka charakterizující chování detektoru tím, že zaznamenává změnu citlivosti v závislosti na změně míry falešně negativních výsledků.



Obr. 7.1: Příklad ROC křivky

Červená čára, která rozděluje graf na dvě poloviny, představuje detekci náhodným výběrem (z toho plyne její poloha, neboť náhodná detekce má stejnou pravděpodobnost zařazení do třídy pozitivní i negativní). Křivka potom představuje zkoumaný detektor. Čím větší je konvexnost teto křivky, tím je daný model účinnější. Snahou je totiž přiblížit modely co nejvíce souřadnici (1,1), která představuje bod optimálního detektoru.

7.2 AUC

Area under the ROC curve (AUC), nebo-li plocha pod křivkou, je mírou určování kvality modelu. Získává se výpočtem velikosti plochy pod křivkou ROC. Čím větší tato plocha je, tím více se blíží bodu (1,1), který představuje ideální detektor. Tato metrika bývá také označována jako diskriminace, tedy schopnost správně rozlišovat negativní a pozitivní instance.

Tyto metriky lze samozřejmě použít i v případě klasifikace do více tříd, nicméně se posuzuje každý model zvlášť, kdy pozitivní stav znamená příslušnost k dané třídě. Patří sem následující metriky:

- TPR (true positive rate) - určuje, kolik instancí do třídy patřící bylo správně klasifikovaných, také bývá označována jako *citlivost* nebo *recall*
- TNR (true negative rate) - určuje, kolik instancí do třídy nepatřící bylo správně klasifikovaných, bývá též označována jako *specificita*
- FPR (false positive rate) - určuje, kolik instancí do třídy nepatřící bylo špatně klasifikovaných jako patřící do třídy
- FNR (false negative rate) - určuje, kolik instancí do třídy patřící bylo špatně klasifikovaných jako nepatřící do třídy

7.3 Výpočet hodnot TPR a TNR

Účinnost VAD metod je vyhodnocena na základě úspěšné detekce segmentů řeči a segmentů pauz.

Jako True positive (TP) je vyhodnocena pouze shoda v identifikaci řeči jak v detekci určitou metodou (reálný VAD), tak i s řečovou aktivitou získanou z manuální segmentace řečového signálu (ideální VAD).

True negative (TN) je vyhodnocen při shodné detekci pauzy v reálném VADu a v ideálním VADu.

	TN	TN		TP		TP	TN	TN	
ideální VAD	0	0	1	1	0	1	0	0	1
reálný VAD	0	0	0	1	1	1	0	0	0

kde 0 = pauza
1 = řeč

Hodnota True positive rate (TPR) je poměr TP k celkovému počtu řečových aktivit získaných z manuální segmentace řečového signálu (ideál VAD).

$$TPR = \frac{TP}{IVP} \quad \text{kde } IVP \text{ – počet 1 v ideálním VADu} \tag{7.1}$$

Hodnota True negative rate (TNR) je poměr TN k celkovému počtu řečových pauz získaných z manuální segmentace řečového signálu (ideál VAD).

$$TNR = \frac{TN}{IVN} \quad \text{kde } IVN \text{ – počet 0 v ideálním VADu} \tag{7.2}$$

TPR a TNR jsou následně vyneseny do grafu a bod, který tyto hodnoty reprezentuje, určuje míru úspěšnosti detekce segmentu řeči a detekce segmentu pauzy. Čím více se bod přiblíží souřadnicím (1,1), tím je detekce řečové aktivity a řečových pauz považována za úspěšnější.

8. Program Matlab

MATLAB je programové prostředí a skriptovací programovací jazyk pro vědeckotechnické numerické výpočty, modelování, návrhy algoritmů, počítačové simulace, analýzu a prezentaci dat, měření a zpracování signálů, návrhy řídicích a komunikačních systémů.

Výpočetní systém MATLAB se během uplynulých let stal celosvětovým standardem v oblasti technických výpočtů a simulací ve sféře vědy, výzkumu, průmyslu i v oblasti vzdělávání.

Název MATLAB vznikl zkrácením slov MATrix LABoratory (volně přeloženo „laboratoř s maticemi“), což odpovídá skutečnosti, že klíčovou datovou strukturou při výpočtech v MATLABu jsou matice. Vlastní programovací jazyk vychází z jazyka Fortran.

MATLAB poskytuje uživatelům nejen mocné grafické a výpočetní nástroje, ale i rozsáhlé specializované knihovny funkcí spolu s výkonným programovacím jazykem čtvrté generace. Knihovny jsou svým rozsahem využitelné prakticky ve všech oblastech lidské činnosti.

Díky své architektuře je MATLAB určen zejména těm, kteří potřebují řešit početně náročné úlohy a přitom nechtějí nebo nemají čas zkoumat matematickou podstatu problémů.

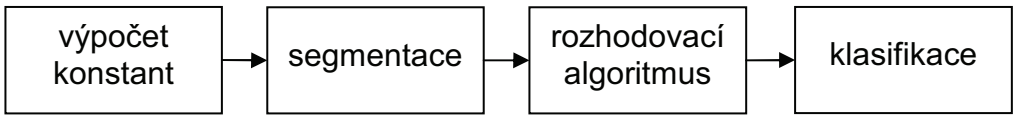
Za nejsilnější stránku MATLABu je považováno mimořádně rychlé výpočetní jádro s optimálními algoritmy, které jsou prověřeny léty provozu na špičkových pracovištích po celém světě. MATLAB byl implementován na všech významných platformách (Windows, Linux, Solaris, Mac).

Nástavbou Matlabu je Simulink – program pro simulaci a modelování dynamických systémů, který využívá algoritmy Matlabu pro numerické řešení především nelineárních diferenciálních rovnic.

9. Použité metody identifikace pauz

9.1 VAD metoda střední hodnoty

Blokové schéma VAD metody střední hodnoty:

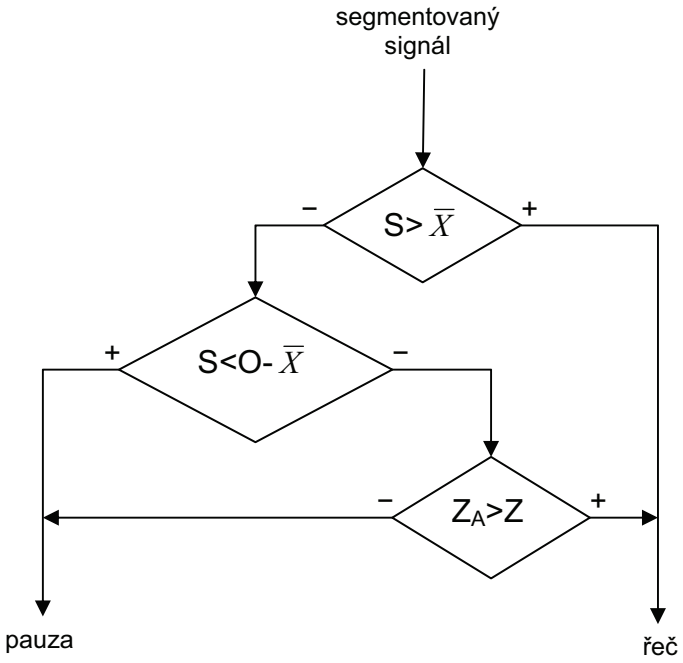


Obr. 9.1: Blokové schéma VAD metody střední hodnoty

Algoritmus detektoru (viz obr. 9.2) v prvním kroku porovná aktuální vzorek signálu *S* se střední hodnotou signálu $\overline{X}$ (9.1). Jestliže je výsledek pozitivní, tzn. aktuální vzorek „*S*“ je větší hodnoty než střední hodnota intenzity signálu $\overline{X}$, může algoritmus s velkou přesností klasifikovat signál jako řeč.

V případě negativního výsledku porovná aktuální vzorek signálu „*S*“ s rozdílem střední hodnoty signálu $\overline{X}$ a jeho směrodatné odchylky *O* (9.2). V případě, že je aktuální vzorek menší než tento rozdíl, klasifikuje signál jako pauzu.

Jestliže se aktuální vzorek nachází v intervalu $<\overline{X}, O-\overline{X}>$, využije znalost o počtu průchodů signálu nulou (9.3), která je spočítána jako průměr s velikostí okna dle vzorkovací frekvence. Nakonec porovná aktuální počet vzorků Z_A se střední hodnotou počtu průchodů signálu nulou získaných využitím průměru *Z* a klasifikuje, zda se jedná o řeč nebo pauzu. [4]



Obr. 9.2: Vývojový diagram rozhodovacího algoritmu VAD metody střední hodnoty

kde:

S	aktuální vzorek signálu
$\overline{X}$	střední hodnota signálu (10.1)
O	směrodatná odchylka signálu (10.2)
Z_A	aktuální počet průchodů signálu nulovou úrovní (10.3)
Z	střední hodnota počtu průchodů signálu nulovou úrovní

9.1.2 Střední hodnota

Nejčastěji používanou charakteristikou střední hodnoty je aritmetický průměr. Aritmetický průměr $\left(\overline{x}\right)$ je nejznámější odhad střední hodnoty, počítá se jako součet všech hodnot vydělených jejich počtem:

$$\overline{\mathrm{x}} = \frac{1}{n}\left(x_1 + x_2 + \ldots + x[n]\right) = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (9.1)$$

Výhodami aritmetického průměru jsou především snadný výpočet a názorný význam.

Nevýhodou je především značná citlivost výsledku na odlehlé hodnoty (např. průměrná mzda) a nevhodnost využití metody u asymetricky rozložených dat.

9.1.3 Směrodatná odchylka

Směrodatná odchylka je v teorii pravděpodobnosti a statistice často používanou mírou statistické disperze. Jedná se o kvadratický průměr odchylek hodnot znaku od jejich aritmetického průměru.

Zhruba řečeno vypovídá o tom, jak moc se od sebe navzájem liší typické případy v souboru zkoumaných čísel. Je-li malá, jsou si prvky souboru většinou navzájem podobné, a naopak velká směrodatná odchylka signalizuje velké vzájemné odlišnosti.

$$o = \left(\frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2 \right)^{\frac{1}{2}} \tag{9.2}$$

kde $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$

9.1.4 Průchod signálu nulovou úrovní

Pokud je signál rozdělený na rámce, lze pro každý z nich určit počet průchodů nulou Z. Tento parametr se využívá k určení znělých a neznělých hlásek. Při vytváření neznělých hlásek se nepoužívají hlasivky, proto se více podobají šumu a mají tedy vyšší Z. Počet průchodů nulou je ale značně náchylný na šum, je proto někdy problém rozlišit neznělou hlásku od šumu.

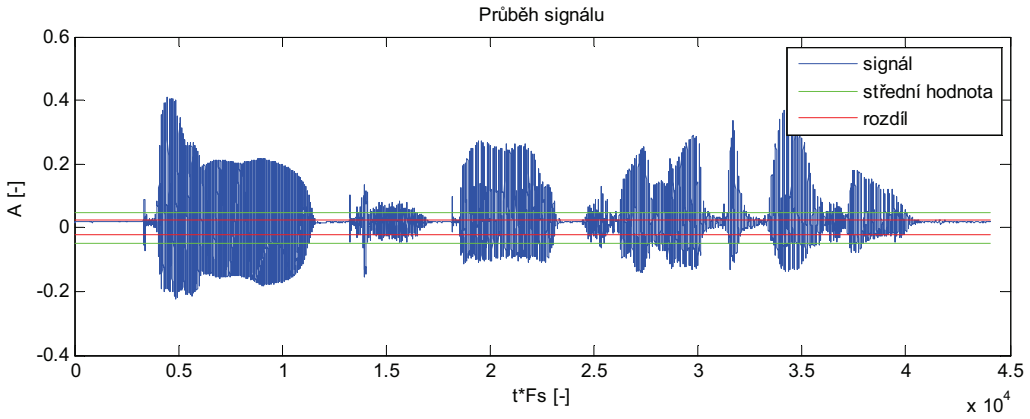
$$Z = \frac{1}{2} \sum_{n=1}^{ramec} \left| signx[n] - signx[n-1] \right| \tag{9.3}$$

kde $\begin{cases} sign\ x[n] = 1 & pro\ x[n] \succ 0 \\ sign\ x[n] = -1 & pro\ x[n] \prec 0 \\ sign\ x[n] = 0 & pro\ x[n] = 0 \end{cases}$

9.1.5 Průběh rozhodovacího algoritmu

Na obr. 9.3 je znázorněna ukázka průběhu nahrávky s vyznačeným průběhem střední hodnoty signálu $\overline{X}$ (9.1) a hodnotou rozdílu střední hodnoty $\overline{X}$ a směrodatné odchylky *O* (9.2).

Rámce, jejichž aktuální střední hodnota je větší než střední hodnota signálu, lze s největší pravděpodobností klasifikovat jako řeč. Naopak rámce, jejichž střední hodnota je menší než rozdíl střední hodnoty a směrodatné odchylky, lze považovat za šum (pauzu).



Obr. 9.3: Ukázka průběhu řečového signálu s vyznačením vypočtených konstant

V dalším kroku se pak určuje, zda je aktuální počet průchodů nulovou úrovní Z_A větší nebo menší než průměrná hodnota počtu průchodů nulovou úrovní Z. Pokud je Z_A větší než Z – jedná se o segment s řečovou aktivitou, v opačném případě se jedná o segment bez řečové aktivity – pauzu.

9.1.6 Vyhodnocení TPR a TNR

Všechny nahrávky z databáze byly testovány VAD metodou střední hodnoty pro segmenty s délkou 32ms a 64ms. Každý tento segment byl vyhodnocen buď jako segment s řečovou aktivitou nebo jako segment bez řečové aktivity – pauza.

Výstupem detekce je binární vektor (1 a 0), kde 1 vyjadřuje segment s řečovou aktivitou a 0 vyjadřuje segment bez řečové aktivity.

Po detekci byl proveden výpočet TP a TN. Následně byly spočítány hodnoty True positive rate (TPR) a True negative rate (TNR).

U této metody se nenastavuje žádný vlastní práh, podle kterého by se posuzoval jednotlivý segment, tudíž jsou přímo spočítány hodnoty TPR a TNR. Hodnoty TPR a TNR jsou následně zapsány do tabulky hodnot pro každou jednotlivou nahrávku. Hodnota TPR a TNR určuje míru úspěšnosti detekce segmentu řeči a detekce segmentu pauzy. Čím více se hodnota přiblíží (1,1), tím je detekce řečové aktivity a řečových pauz považována za úspěšnější.

Tab. 9.1: Hodnoty TPR a TNR pro 32ms segment:

Název souboru	TPR	TNR
Alessio.wav	0,5756	0,9497
Antonio.wav	0,5353	0,9755
Carmine.wav	0,6148	0,9954
Franco.wav	0,5385	0,7096
Gerry.wav	0,5532	0,4446
Marco_magliuolo.wav	0,5768	0,8909
Marco_riccio.wav	0,5836	0,9903
Paolo.wav	0,5543	0,8987
Tonyf.wav	0,5374	0,6300

Z tabulky 9.1 je patrné, že nejhůře dopadly nahrávky se vzorkovací frekvencí F_{vz} = 16kHz (Franco.wav, Gerry.wav a Tonyf.wav), což je způsobeno nižší vzorkovací frekvencí F_{vz}.

Větší úspěšnost detekce je u segmentů bez řečové aktivity – TNR.

Tab. 9.2: Hodnoty TPR a TNR pro 64ms segment:

Název souboru	TPR	TNR
Alessio.wav	0,6054	0,9773
Antonio.wav	0,5457	0,9869
Carmine.wav	0,6459	0,9971
Franco.wav	0,5670	0,8675
Gerry.wav	0,5781	0,6174
Marco_magliuolo.wav	0,5945	0,9225
Marco_riccio.wav	0,5911	0,9957
Paolo.wav	0,5699	0,9045
Tonyf.wav	0,5720	0,7451

Zdvojnásobením délky segmentu z délky 32ms na 64ms došlo u všech nahrávek ke zlepšení detekce segmentů s řečovou aktivitou a segmentů bez řečové aktivity – pauz. Stejně jako u segmentů s délkou 32ms byly i zde výrazně horší výsledky detekce u nahrávek s nižší vzorkovací frekvencí F_{vz}.

9.2 VAD metoda FFT

9.2.1 Rychlá Fourierova transformace

Diskrétní Fourierova transformace našla velké uplatnění zejména s rozvojem výpočetní techniky. Součástí řady přístrojů jsou jednoúčelové procesory realizující tuto transformaci. Její hlavní rozvoj nastal po roce 1965, kdy J.W. Cooley a J.W. Tukey popsali velmi efektivní algoritmus výpočtu DFT, tzv. rychlou Fourierovu transformaci (FFT - Fast Fourier Transform). Díky tomuto algoritmu se stala diskrétní Fourierova transformace nejrozšířenějším prostředkem pro numerický výpočet Fourierovy transformace. Algoritmus FFT je také implementován ve všech nejrozšířenějších matematických programech jako je např. Matlab, GNU Octave, Mathcad, Mathematica, Maple, atd.

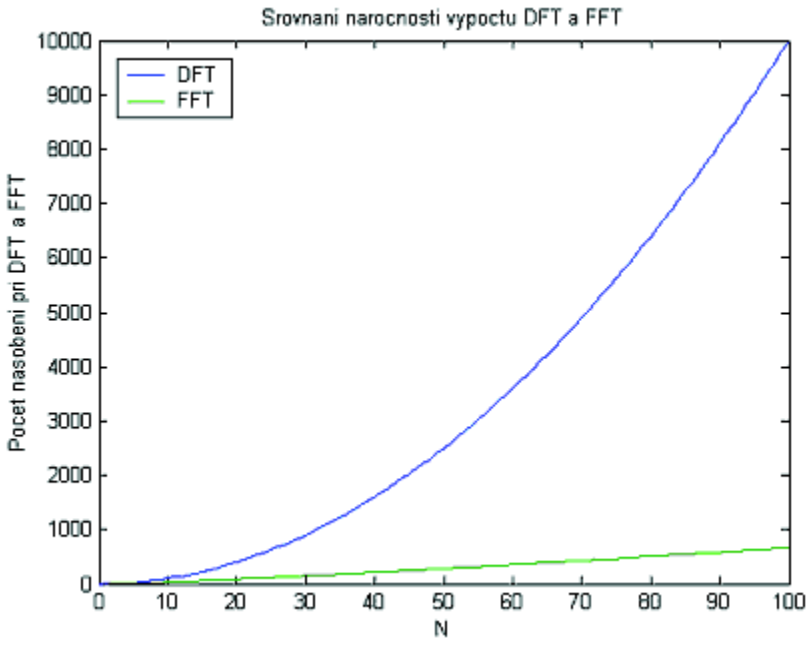
Pravda je však taková, že tato technika byla objevena několik desetiletí před rokem 1965. Například Německy matematik Karl Friedrich Gauss ji popsal o více jak sto let dříve. Jeho práce však upadla v zapomnění, protože v jeho době neexistovala hlavní věc, pro kterou byla FFT navržena - digitální počítač.

FFT je rychlý algoritmus pro počítání diskrétní Fourierovy transformace (DFT) a její inverze. FFT je velice důležitá pro široký okruh aplikací, kde se pracuje s digitálním zpracováním signálu a řešením parciálních diferenciálních rovnic algoritmy, které se využívají pro rychlé násobení čísel. DFT má operační náročnost algoritmu $O(N^2)$ operací, oproti tomu FFT má $O(N \log N)$ operací.

Pro srovnání náročnosti výpočtu DFT a FFT použijme následující rovnici:

$$\frac{N^2}{N\log_2(N)} = \frac{N}{\log_2(N)} = \frac{2^p}{p} \qquad [8] \qquad (9.4)$$

Např. pro p=10 je N=1024 je výpočet FFT 102,4 krát rychlejší než výpočet DFT viz obr. 10.4



[9]

Obr. 9.4: Srovnání náročnosti výpočtu DFT a FFT

Tři hlavní problémy aplikace Fourierova teorému při analýze řeči:

- řečové zvuky jsou kvaziperiodické - proto používáme okno (jeho délka se stává periodou)
- digitální analýza užívá diskrétní signál - transformace DFT (Discrete Fourier Transform)

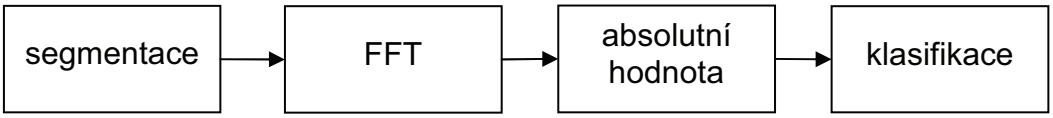
- složité výpočty - algoritmus DFT nazvaný FFT (Fast Fourier Transform)

Podstata FFT spočívá ve spočítání korelace sinusových a kosinusových složek se složenou vlnou. Vysoká korelace znamená, že složená vlna a daná sinusovka stoupají a klesají spolu.

9.2.2 Postup výpočtu

Nejznámějším postupem pro vyjádření vlastností řečového signálu je jeho převod z časové do kmitočtové oblasti. Tímto převodem lze získat mnohem více charakteristických vlastností popisujících zpracovávaný úsek, na jejichž základě je možné rozhodnout o jeho charakteru. Pro tento účel slouží doposud nepřekonaná Fourierova transformace (FT), resp. její rychlejší výpočetní varianta FFT (Fast Fourier Transform). Po sobě jdoucí, nebo překrývající se časové úseky jsou transformovány do frekvenční oblasti. Překrývání úseků aplikujeme z důvodu minimalizace spektrálních nespojitostí na začátku a konci každého časového rámce. Při segmentaci řečového signálu se nejčastěji používá Hammingovo okno.

Blokové schéma použité VAD metody FFT:



Obr. 9.5: Blokové schéma VAD metody FFT

Nahrávky byly nejprve nasegmentovány a potom na ně byla použita rychlá Fourierova transformace (FFT). Po použití FFT je v segmentech komplexní posloupnost. Tato komplexní posloupnost byla absolutní hodnotou převedena na posloupnost reálnou. Pro další výpočet této metody byl použit vzorec uvedený v rovnici (9.5).

$$H = \frac{1}{L}\log\frac{p\left(X \mid \hat{\Theta}, H_S\right)}{p\left(X \mid H_N\right)} = \frac{1}{L}\sum_{k=0}^{L-1}\left\{\frac{|X_k|^2}{\lambda_N(k)} - \log\frac{|X_k|^2}{\lambda_N(k)} - 1\right\} \underset{H_N}{\overset{H_S}{\gtrless}} \eta \quad [7] \qquad (9.5)$$

kde

L	délka segmentu
Θ	{λ(k) : k = 0,…L-1}
X	rušený řečový signál
H_S	segment řeči
H_N	segment šumu
k	pořadí segmentu
X_k	segment rušeného řečového signálu
$\lambda_N(k)$	rozptyl šumu - referenční šumový segment
η	vlastní práh – řeč / pauza

Referenčním šumovým segmentem se rozumí segment řečové nahrávky, který obsahuje pauzu (šum). Tento šumový segment je nadále považován za referenční. Je tedy nutno v řečové nahrávce určit segment, který je šumový. Musí se provést odhadem a nebo jako v tomto případě segment, který byl identifikován na základě manuální segmentace řečového signálu.

Tato metoda je založena na porovnání výsledné hodnoty s hodnotou vlastního prahu η. Pokud je hodnota segmentu větší než hodnota vlastního prahu η, jedná se o segment s řečovou aktivitou, naopak pokud je hodnota segmentu menší než hodnota vlastního prahu η, jedná se o segment bez řečové aktivity – pauzu.

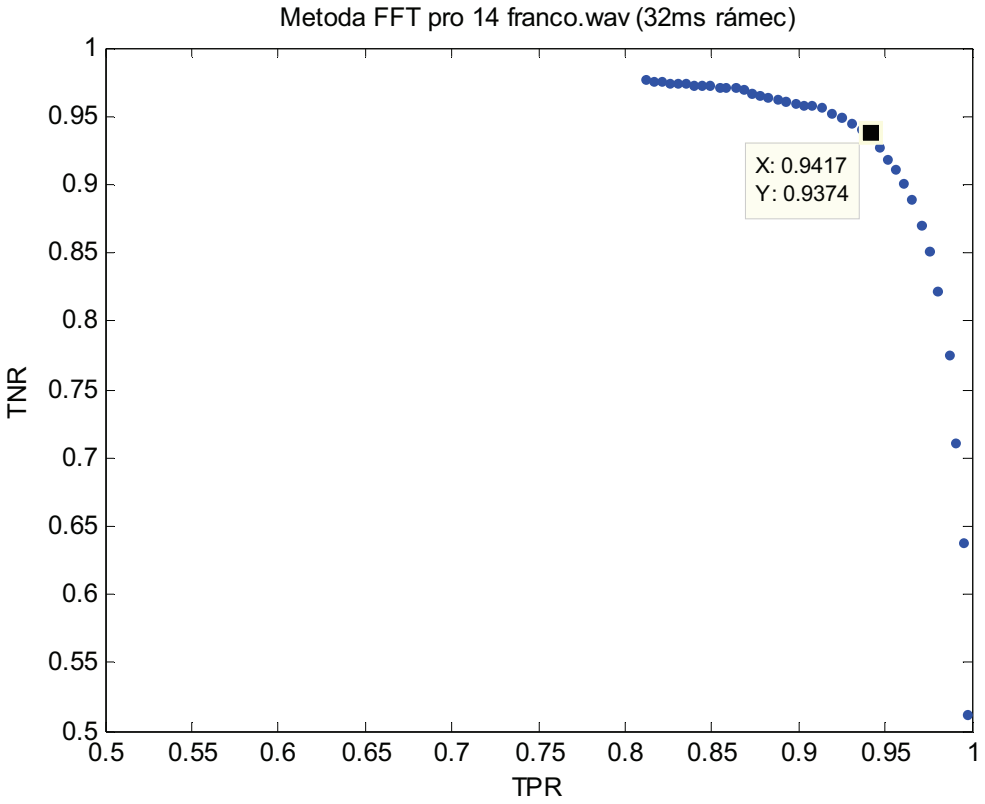
Problémem je určení vlastního prahu, kdy tato hodnota může nabývat různých hodnot a zásadně tím ovlivňuje výsledek této metody. Pro správné určení hodnoty vlastního prahu každé řečové nahrávky bylo použito 40 hodnot s konstantním zvyšováním hodnoty vlastního prahu.

Pro každou hodnotu vlastního prahu byly vypočítány hodnoty TP a TN, z nichž se dále vypočítaly hodnoty TPR a TNR. Všech 40 hodnot TPR a TNR pro každou řečovou nahrávku bylo následně vyneseno do grafu a z něj

odečtena nejlepší hodnota TPR a TNR. Nejlepší hodnota TPR a TNR se nachází co nejblíže úhlopříčce a souřadnicím (1;1).

9.2.3 Postup výpočtu pro segmenty délky 32ms

Jako příklad tohoto postupu je na obr. 9.6 zobrazena ROC křivka pro řečovou nahrávku franco.wav o délce segmentu 32ms. Nejlepší hodnota TPR a TNR byla vyhodnocena na souřadnicích (0,9417 ; 0,9374).



Obr. 9.6: ROC křivka metody FFT pro framco.wav (32ms rámec)

V tabulce 9.3 je zaznamenáno všech 40 hodnot vlastního prahu η. Souřadnice s nejlepší hodnotou TPR a TNR podle tabulky odpovídá vlastnímu prahu η úrovni 35 000. V případě použití tohoto vlastního prahu η pro detekci řečové aktivity této řečové nahrávky (franco.wav) lze dosáhnout úspěšnosti detekce segmentu s řečovou aktivitou v 94,17% případů a v případě detekce segmentu bez řečové aktivity lze dosáhnout úspěšnosti v 93,74% případů.

Tab. 9.3: 40 hodnot TPR a TNR pro franco.wav (32ms rámec) s vlastním prahem η:

TPR	TNR	η
1,0000	0,0005	0
0,9998	0,2593	2500
0,9971	0,5120	5000
0,9950	0,6382	7500
0,9901	0,7112	10000
0,9863	0,7750	12500
0,9801	0,8226	15000
0,9754	0,8516	17500
0,9702	0,8704	20000
0,9647	0,8901	22500
0,9604	0,9017	25000
0,9560	0,9113	27500
0,9513	0,9195	30000
0,9467	0,9276	32500
0,9417	**0,9374**	**35000**
0,9365	0,9408	37500
0,9306	0,9456	40000
0,9244	0,9491	42500
0,9191	0,9524	45000
0,9127	0,9567	47500
0,9077	0,9586	50000
0,9027	0,9590	52500
0,8981	0,9600	55000
0,8925	0,9615	57500
0,8871	0,9624	60000
0,8823	0,9637	62500
0,8776	0,9656	65000
0,8730	0,9673	67500
0,8681	0,9694	70000
0,8635	0,9708	72500
0,8582	0,9718	75000
0,8539	0,9720	77500
0,8489	0,9728	80000
0,8442	0,9728	82500
0,8393	0,9734	85000
0,8350	0,9739	87500
0,8306	0,9744	90000
0,8258	0,9749	92500
0,8210	0,9753	95000
0,8164	0,9758	97500
0,8118	0,9772	100000

Pro další řečové nahrávky z databáze už nebudou uváděny tabulky se všemi vypočítanými hodnotami, ale je pouze v tabulce uvedena nejlepší hodnota TPR a TNR a k těm příslušná hodnota vlastního prahu η viz tab. 9.4.

Tab. 9.4: Nejlepší vypočtené hodnoty TPR a TNR pro všechny řečové nahrávky s rámcem 32ms metodou FFT s určením vlastního prahu η:

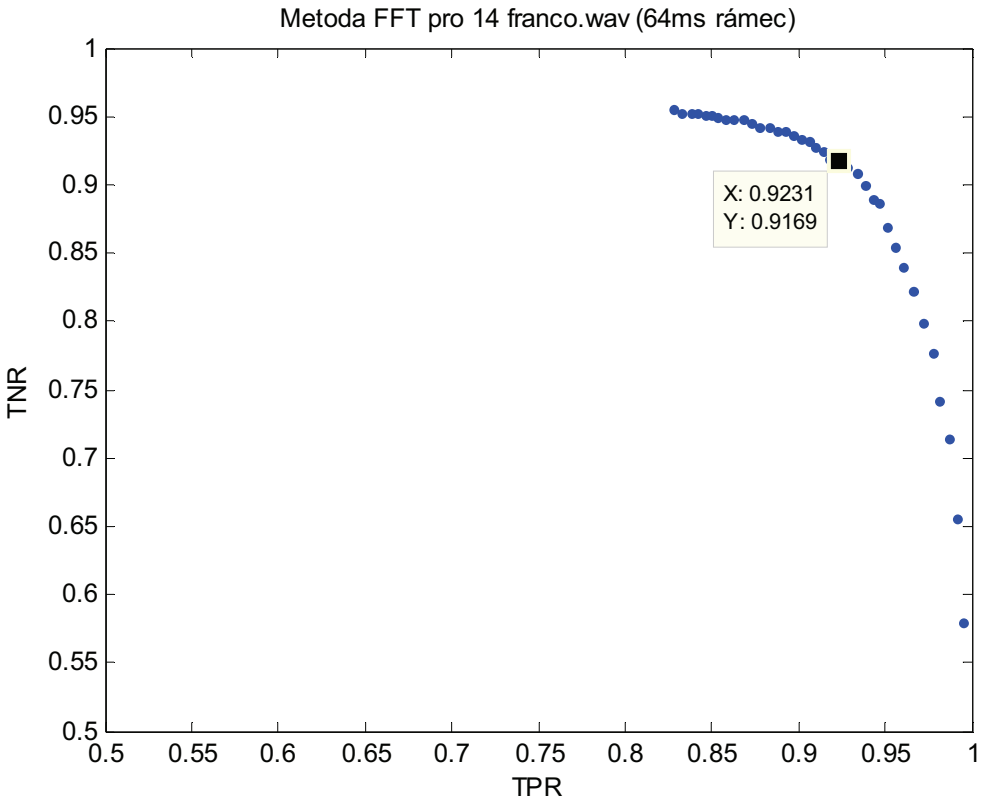
Název souboru	TPR	TNR	η
Alessio.wav	0,6653	0,6875	200000
Antonio.wav	0,8605	0,8610	14000
Carmine.wav	0,8926	0,8946	10000
Franco.wav	0,9417	0,9374	35000
Gerry.wav	0,9048	0,9110	50000
Marco_magliuolo.wav	0,6550	0,6371	14000
Marco_riccio.wav	0,7308	0,7305	13500
Paolo.wav	0,7648	0,7718	6750
Tonyf.wav	0,8387	0,8427	400000

Graf s průběhem ROC křivky pro každou další řečovou nahrávku je umístěn v příloze na konci této diplomové práce. V každém grafu je vyznačena nejlepší hodnota TPR a TNR, které odpovídají údajům ve výše uvedené tabulce. Pro každou řečovou nahrávku je v tabulce uvedena i odpovídající hodnota vlastního prahu η.

Zhoršené výsledky u některých řečových nahrávek mohou být způsobeny nevhodným výběrem referenčního šumového segmentu λ.

9.2.4 Postup výpočtu pro segmenty délky 64ms

Jako příklad tohoto postupu je na obr. 9.7 zobrazena ROC křivka pro řečovou nahrávku franco.wav o délce segmentu 64ms. Nejlepší hodnota TPR a TNR byla vyhodnocena na souřadnicích (0,9231 ; 0,9169).



Obr. 9.7: ROC křivka metody FFT pro framco.wav (64ms rámec)

V tabulce 9.5 je zaznamenáno všech 40 hodnot vlastního prahu η. Souřadnice s nejlepší hodnotou TPR a TNR podle tabulky odpovídá vlastnímu prahu η úrovni 190 000. V případě použití tohoto vlastního prahu η pro detekci řečové aktivity této řečové nahrávky (franco.wav) lze dosáhnout úspěšnosti detekce segmentu s řečovou aktivitou v 92,31% případů a v případě detekce segmentu bez řečové aktivity lze dosáhnout úspěšnosti v 91,69% případů.

Tab. 9.5: 40 hodnot TPR a TNR pro franco.wav (64ms rámec) s vlastním prahem η:

TPR	TNR	eta
0,9996	0,1480	10000
0,9989	0,3590	20000
0,9971	0,4889	30000
0,9948	0,5804	40000
0,9911	0,6554	50000
0,9864	0,7146	60000
0,9808	0,7425	70000
0,9771	0,7767	80000
0,9718	0,7988	90000
0,9663	0,8228	100000
0,9608	0,8398	110000
0,9558	0,8541	120000
0,9515	0,8694	130000
0,9468	0,8862	140000
0,9431	0,8900	150000
0,9379	0,8993	160000
0,9337	0,9086	170000
0,9277	0,9128	180000
0,9231	**0,9169**	**190000**
0,9174	0,9191	200000
0,9137	0,9246	210000
0,9100	0,9274	220000
0,9056	0,9327	230000
0,9018	0,9341	240000
0,8968	0,9363	250000
0,8927	0,9386	260000
0,8875	0,9396	270000
0,8831	0,9421	280000
0,8775	0,9421	290000
0,8732	0,9449	300000
0,8676	0,9477	310000
0,8624	0,9478	320000
0,8582	0,9487	330000
0,8534	0,9499	340000
0,8494	0,9512	350000
0,8458	0,9512	360000
0,8419	0,9521	370000
0,8378	0,9525	380000
0,8327	0,9526	390000
0,8281	0,9556	400000

Pro další řečové nahrávky z databáze už nebudou uváděny tabulky se všemi vypočítanými hodnotami, ale je pouze v tabulce uvedena nejlepší hodnota TPR a TNR a k těm příslušná hodnota vlastního prahu η.

Tab. 9.6: Nejlepší vypočtené hodnoty TPR a TNR pro všechny řečové nahrávky s rámcem 64ms metodou FFT s určením vlastního prahu η:

Název souboru	TPR	TNR	eta
Alessio.wav	0,7556	0,7579	22000
Antonio.wav	0,8119	0,8126	35000
Carmine.wav	0,8856	0,8669	32500
Franco.wav	0,9231	0,9169	190000
Gerry.wav	0,8955	0,8944	325000
Marco_magliuolo.wav	0,6381	0,6405	69000
Marco_riccio.wav	0,8744	0,8692	22000
Paolo.wav	0,7766	0,7869	23000
Tonyf.wav	0,8447	0,8494	220000

Graf s průběhem ROC křivky pro každou další řečovou nahrávku je umístěn v příloze na konci této diplomové práce. V každém grafu je vyznačena nejlepší hodnota TPR a TNR, které odpovídají údajům ve výše uvedené tabulce. Pro každou řečovou nahrávku je v tabulce uvedena i odpovídající hodnota vlastního prahu η.

Zhoršené výsledky u některých řečových nahrávek mohou být způsobeny nevhodným výběrem referenčního šumového segmentu λ.

10. Závěr

Cílem diplomové práce bylo najít efektivní metody, které by dokázaly identifikovat pauzy bez řečové aktivity.

Diplomová práce se věnuje dvěma metodám pro identifikaci pauzy. První metodou je metoda střední hodnoty a druhou je metoda s použitím rychlé Fourierovi transformace (FFT).

Obě dvě metody identifikace pauz v rušeném řečovém signálu byly implementovány v MATLABu. Soubory s naprogramovanými metodami jsou součástí přiloženého CD.

Před samotnou analýzou metod je potřeba nastavit všechny vstupní parametry, jejichž hodnotu lze měnit. U obou metod je nutné nastavit délku segmentu (bylo použito délky 32ms a 64ms), u metody FFT je nutné navíc nastavit referenční rozptyl šumu a vlastní práh metody.

Toto nastavení je založeno především na testování úspěšnosti nalezení vlastního prahu a je mu nutné věnovat velkou pozornost, neboť nastavení tohoto parametru zásadně ovlivňuje kvalitu identifikace.

Analýza metod byla provedena na devíti řečových nahrávkách. K vyhodnocení analýzy metody FFT byly použity ROC křivky s vyobrazením 40 hodnot TPR (True Positive Rate) a TNR (True Negative Rate) - výsledné grafy jsou v přílohách diplomové práce.

Výsledky metody FFT pro každou řečovou nahrávku byly vyneseny do grafů s ROC křivkami a na jejich základě byl vyhodnocen nejlepší vlastní práh pro danou řečovou nahrávku. Pro řečovou nahrávku franco.wav byla pro délku segmentu 32ms a 64ms vytvořena tabulka, ve které byly zaznamenány hodnoty TPR, TNR a vlastního prahu – nejlepší vlastní práh je zvýrazněn.

Úspěšnost klasifikace u metody FFT je do jisté míry dána i správným počátečním zadáním segmentu s rozptylem šumu – referenčním šumem.

U metody střední hodnoty se grafy s ROC křivkou nemusely dělat, protože se nemusel hledat další vhodný parametr pro nejlepší nastavení vstupních hodnot metody. U této metody byla zaznamenána pouze hodnota TPR a TNR.

Při porovnání úspěšnosti klasifikace pro jednotlivé délky segmentace je vidět, že při použití delšího segmentu je úspěšnost klasifikace vyšší.

Při vyhodnocení úspěšnosti metod identifikovat pauzu je zřejmé, že metoda FFT dosáhla poměrně dobrých výsledků, naopak metodu střední hodnoty nedoporučuji pro identifikaci pauz v rušeném řečovém signálu používat.

11. Literatura

[1] PSUTKA, Josef. 1995. *Komunikace s počítačem mluvenou řečí.* Praha : Academia Praha, 1995.

[2] PALKOVÁ, Zdena. *Fonetika a fonologie češtin*y. Praha : Karolinum, 1997. 366 s.

[3] KRČMOVÁ, Marie. *Fonetika a fonologie. Zvuková stavba češtiny*. Brno : UJEP, 1990.

[4] RUSZ Jan, ČMEJLA Roman. *Akustická analýza intenzity a rychlosti řeči u Parkinsonovy nemoci.* Praha : Humusoft , 2008.

[5] http://amber.feld.cvut.cz/vyu/zzs/zzs9/c8.htm

[6] PSUTKA, Josef, a další. 2006. *Mluvíme s počítačem česky.* Praha : Academia Praha, 2006.

[7] SOHN, Jongseo, SUNG, Wonyong. *A voice activity detector employing soft decision based noise spectrum adaptation.* Seoul: Seoul National University, 1998.

[8] FRAZIER, M. W. Introduction to wavelets through linear Algebra. Springer, 1998.

[9] MARŠÁLEK, Leoš, SKAPA, Jan. Diskrétní transformace. 2003.

[10] STEJSKAL, Vojtěch. Automatická segmentace řeči a identifikace pauz. VUT Brno, 2006.

[11] http://www.p007.webpark.cz/se/

[12] SOVKA, P., POLLÁK, P.: Vybrané metody číslicového zpracování signálů. *ČVUT,* Praha 2001.

[13] KOČÍ, Miloslav. 2010. Rozpoznávání slov diskrétního diktátu. Univerzita Pardubice: Fakulta elektrotechniky a informatiky, 2010.

[14] HORÁK, Petr. Modelování suprasegmentálních rysů mluvené češtiny pomocí lineární predikce. ČVUT Praha: Fakulta elektrotechnická, 2002

12. Seznam použitých zkratek a symbolů

λ_N	šumový segment - referenční
η	vlastní práh – řeč / pauza
AUC	Area under the ROC curve
DFT	diskrétní Fourierova transformace
E	krátkodobá energie
FFT	Fast Fourier Transform – rychlá Fourierova transformace
f_{vz}	vzorkovací frekvence
f_0	základní tón lidského hlasu
H	hodnota výpočtu segmentu VAD metodou FFT
H_S	hodnota pro určení segmentu jako řeči VAD metodou FFT
H_N	hodnota pro určení segmentu jako pauzy VAD metodou FFT
idealVAD	ručně značené pauzy
L	délka rámce (segmentu)
M	krátkodobá intenzita
N	celkový počet segmentů
O	směrodatná odchylka
R	autokorelční funkce
realVAD	detekované pauzy
ROC	Receiver Operating Characteristic
T	perioda vzorkování
TP	True Positive
TPR	True Positive Rate
TN	True Negative
TNR	True Negative Rate
VAD	Voice Activity Detector - detektor řečové aktivity
x[n]	řečový signál
$\overline{X}$	střední hodnota
Z	průchod signálu nulovou úrovní
Z_A	aktuální hodnota průchodu signálu nulovou úrovní

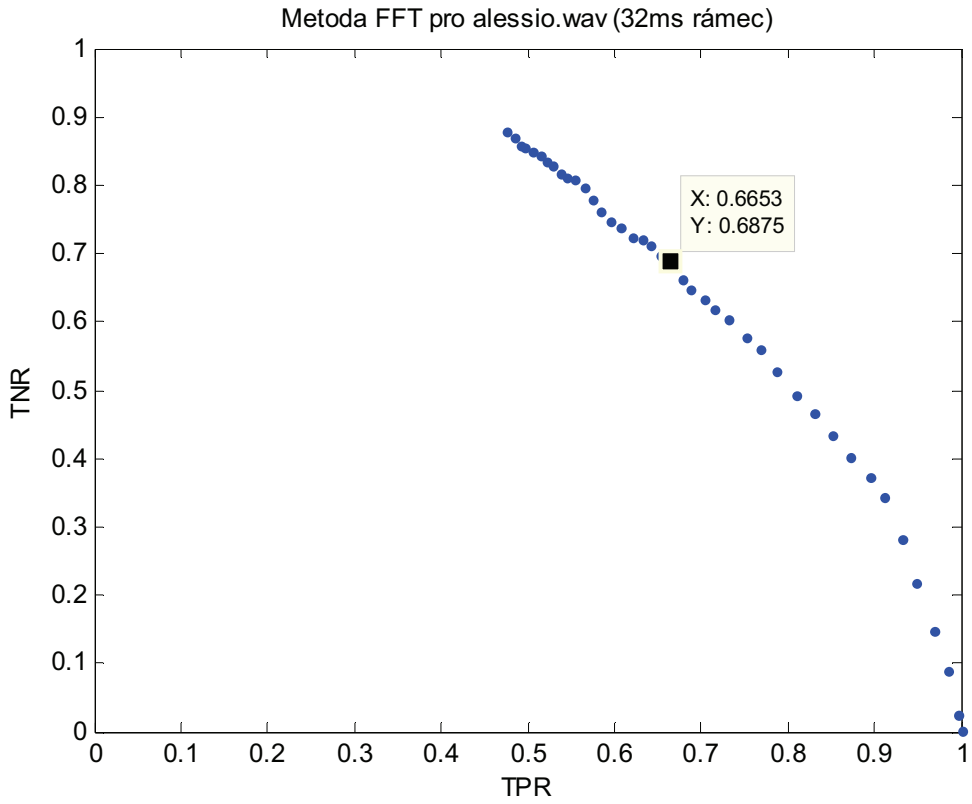
13. Seznam příloh

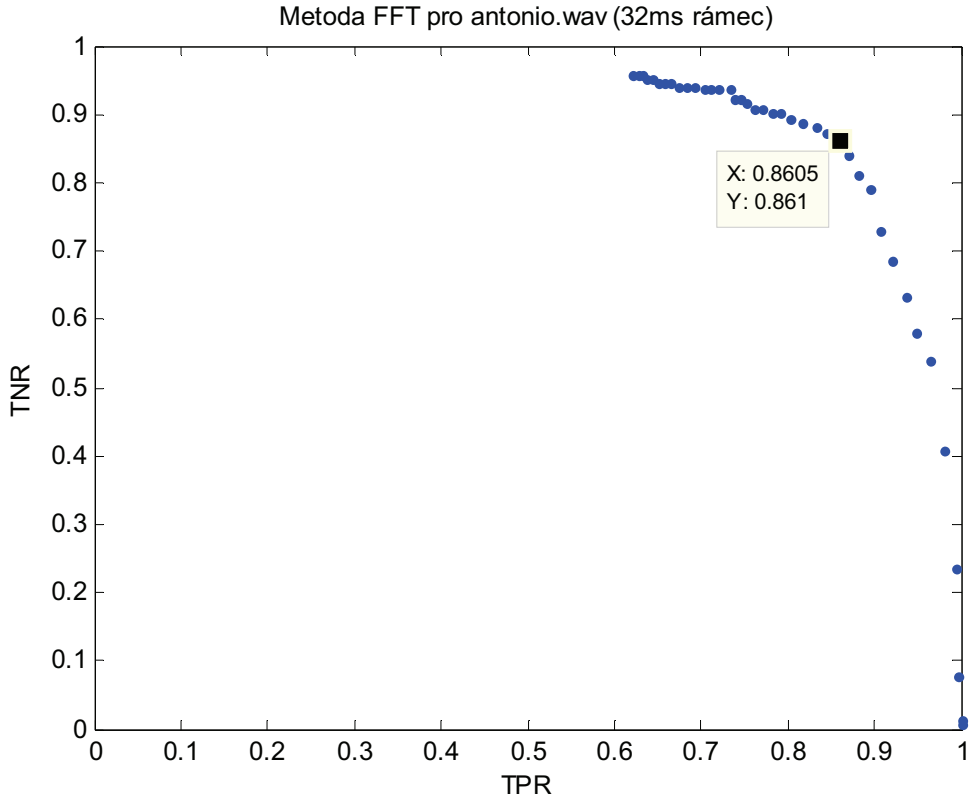
A ROC grafy metody FFT pro segment délky 32ms

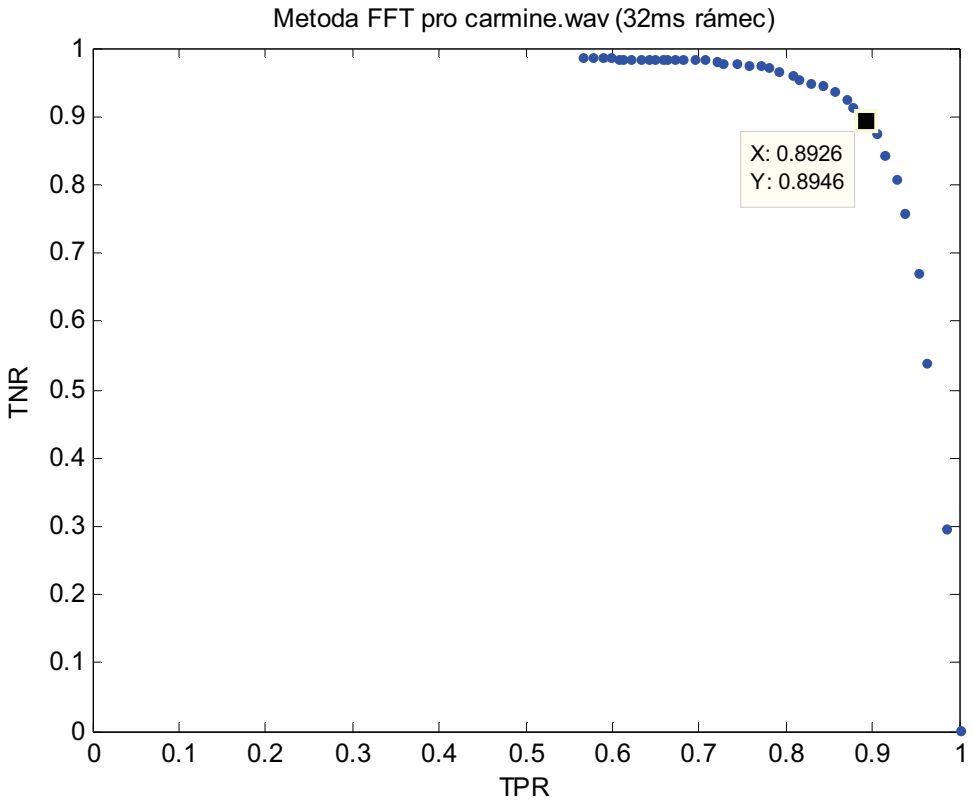
B ROC grafy metody FFT pro segment délky 64ms

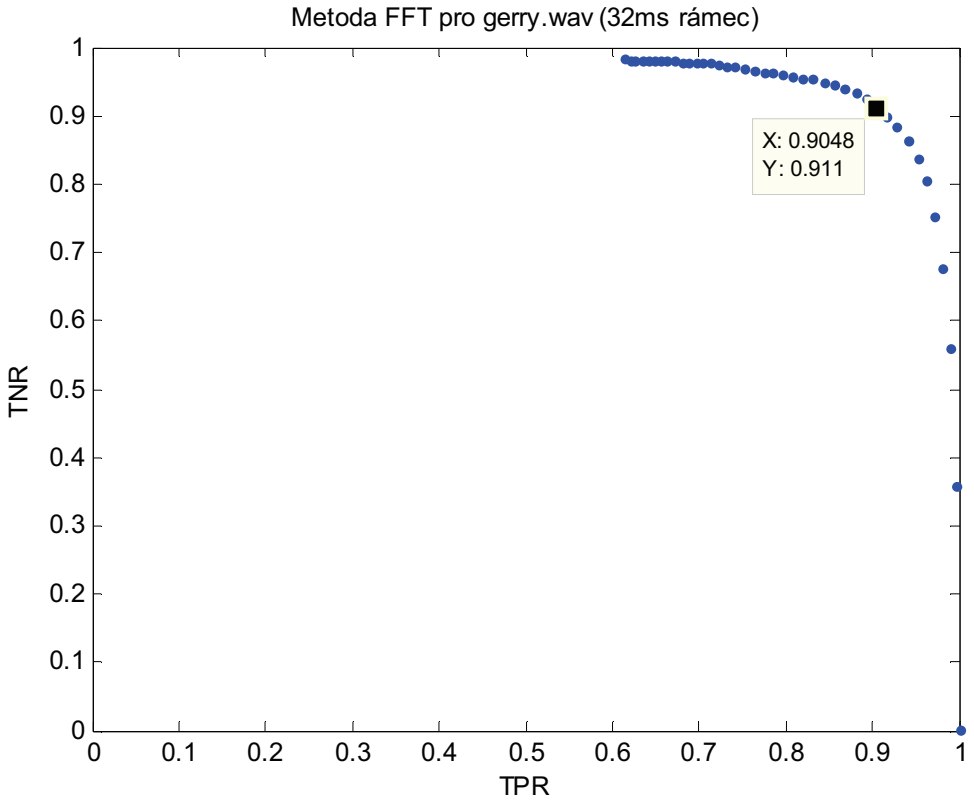
C obsah přiloženého CD

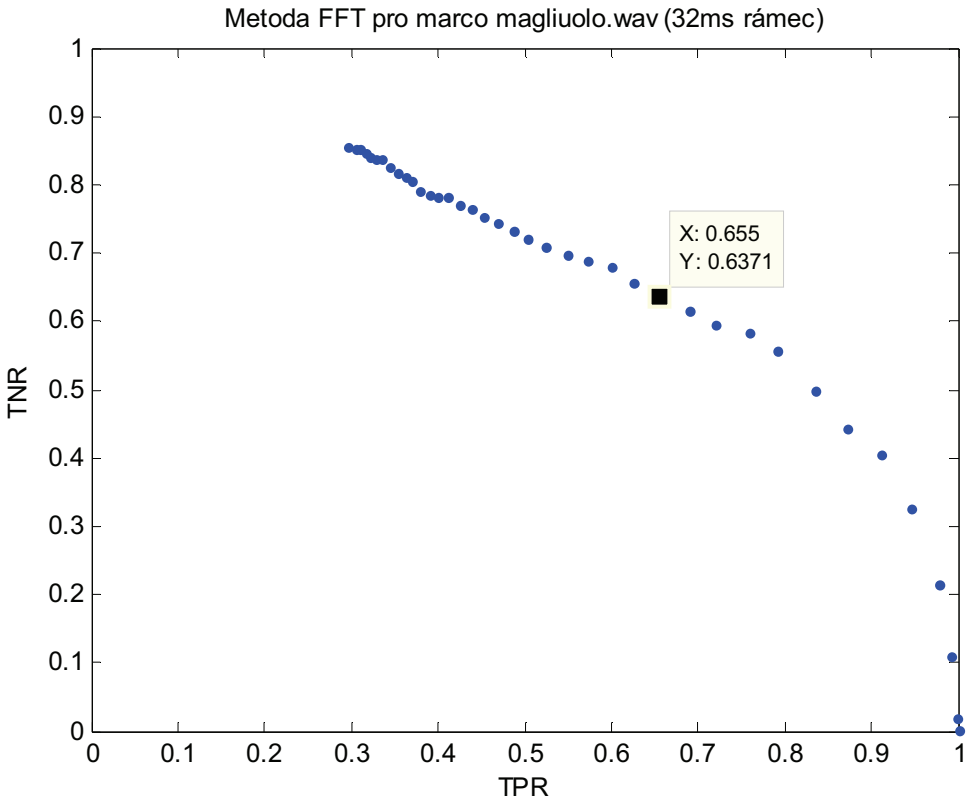
A ROC grafy metody FFT pro segment délky 32ms

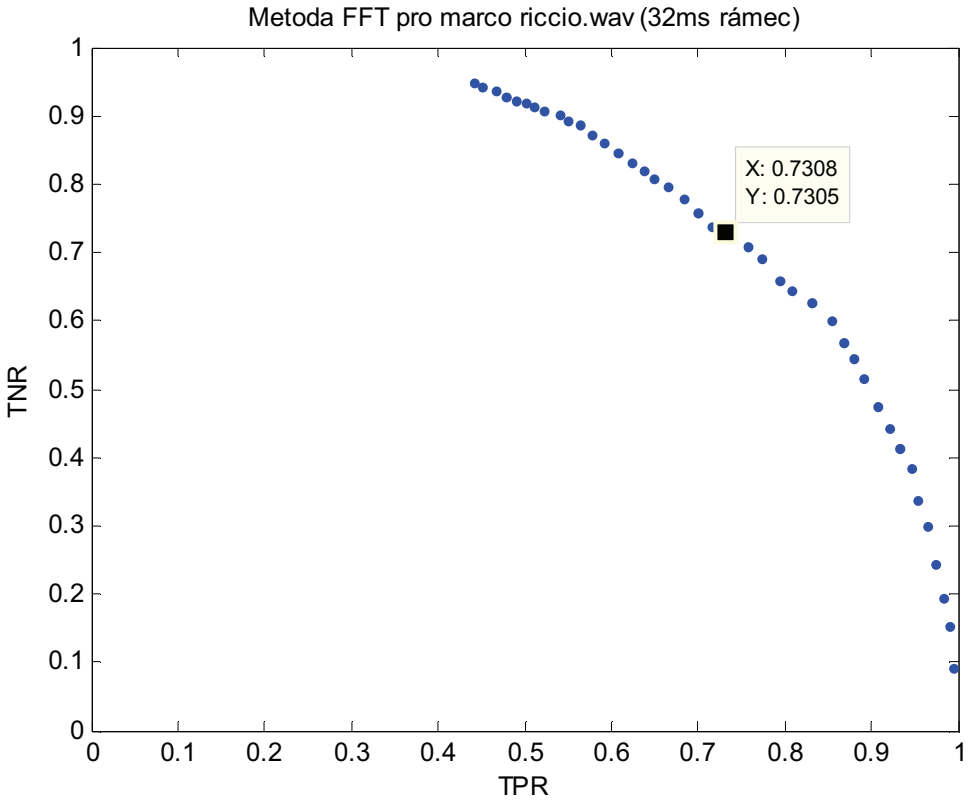


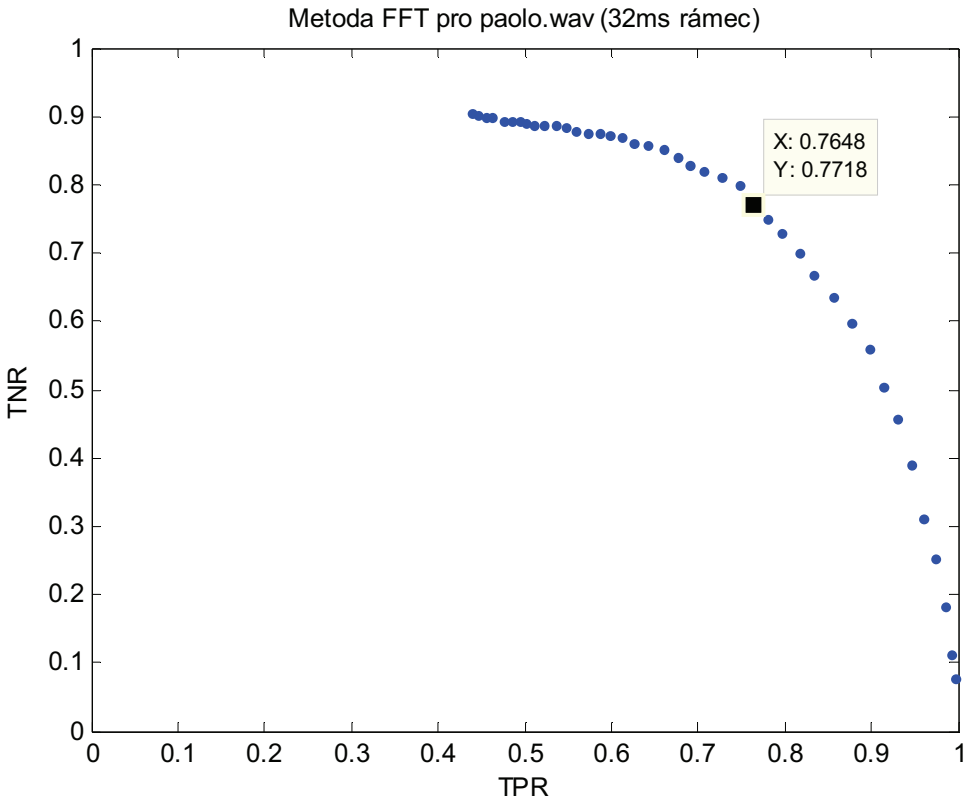


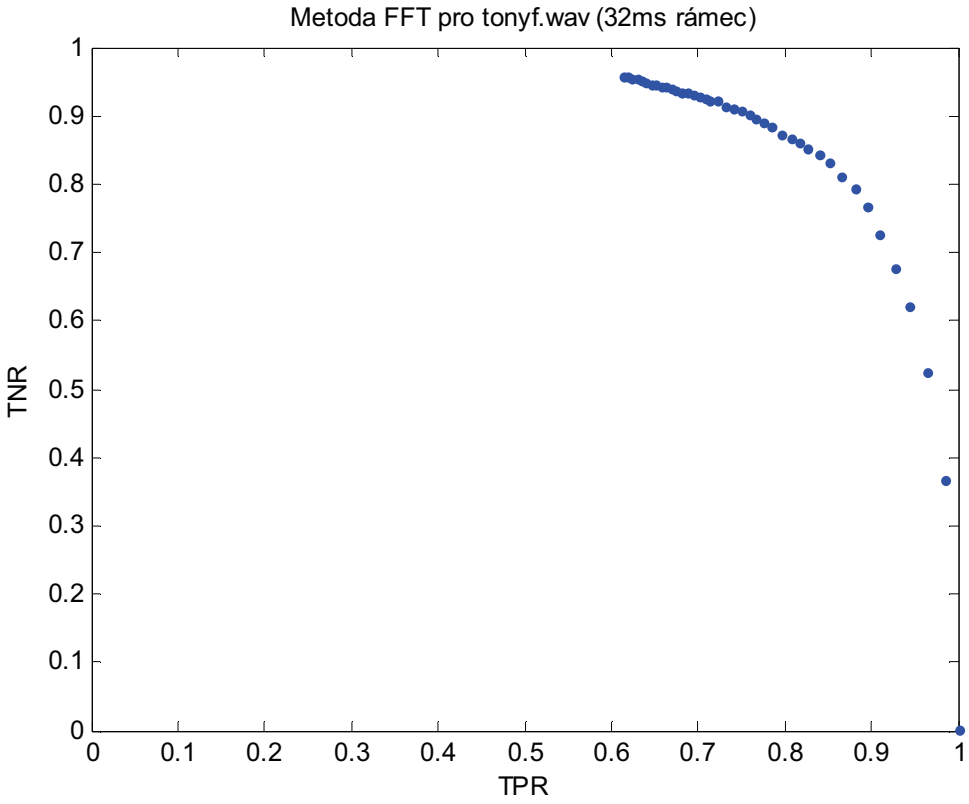

Metoda FFT pro carmine.wav (32ms rámec)
X: 0.8926
Y: 0.8946
TNR
TPR
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


Metoda FFT pro gerry.wav (32ms rámec)
X: 0.9048
Y: 0.911
TNR
TPR
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

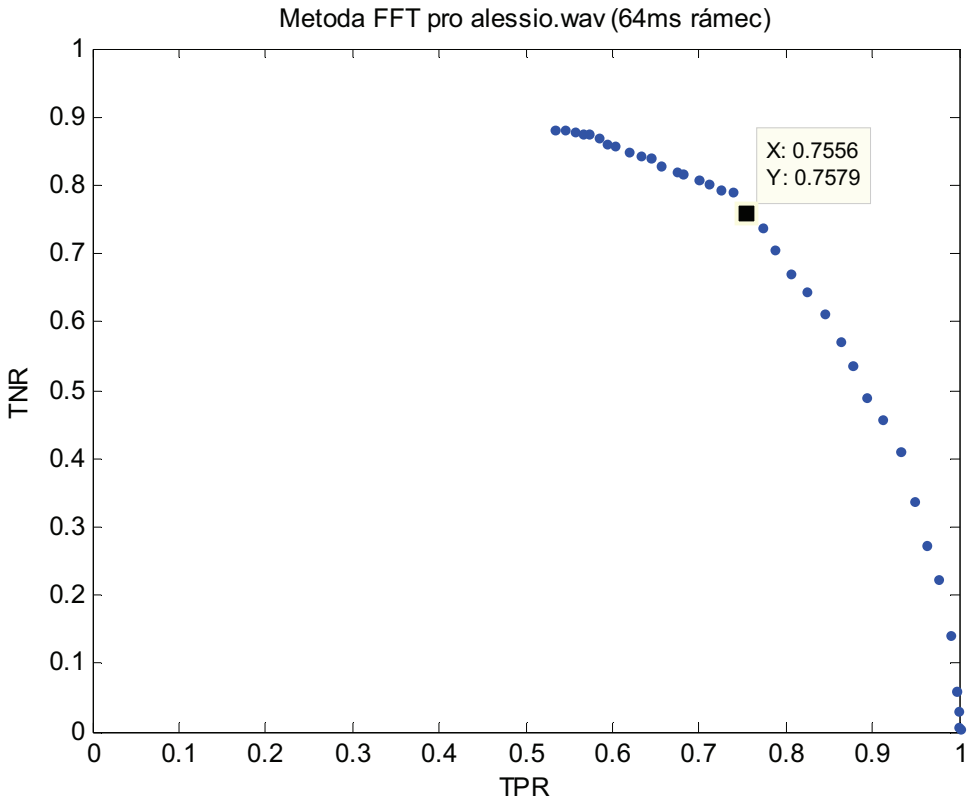

Metoda FFT pro marco magliuolo.wav (32ms rámec)
X: 0.655
Y: 0.6371
TNR
TPR
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

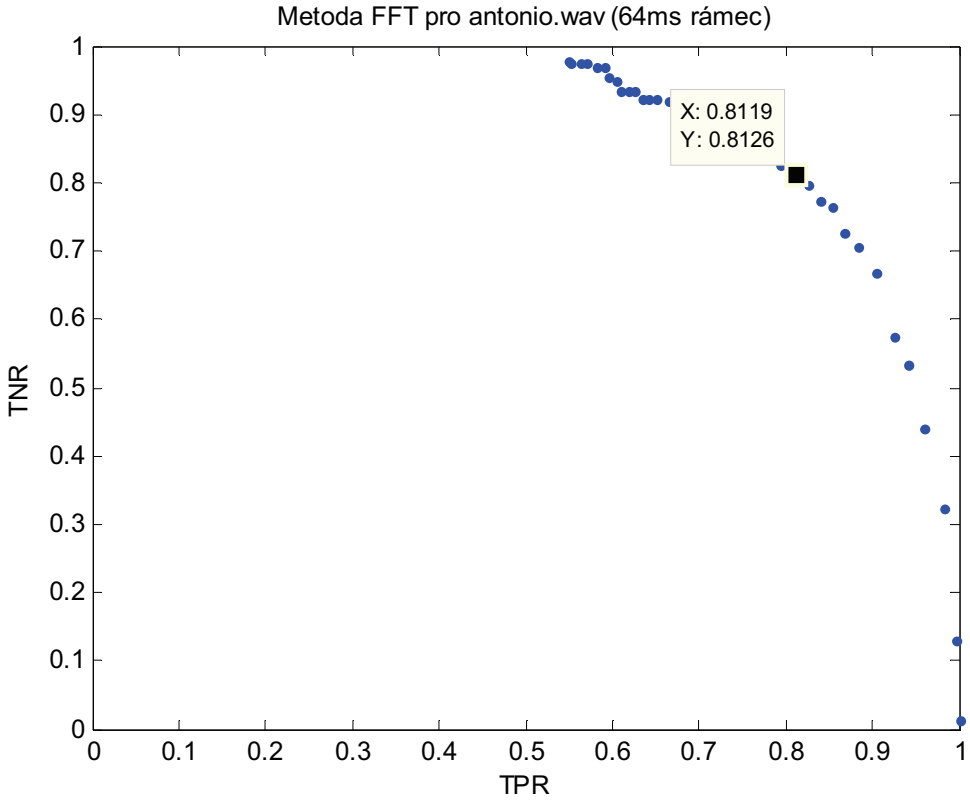

Metoda FFT pro marco riccio.wav (32ms rámec)
X: 0.7308
Y: 0.7305
TNR
TPR
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

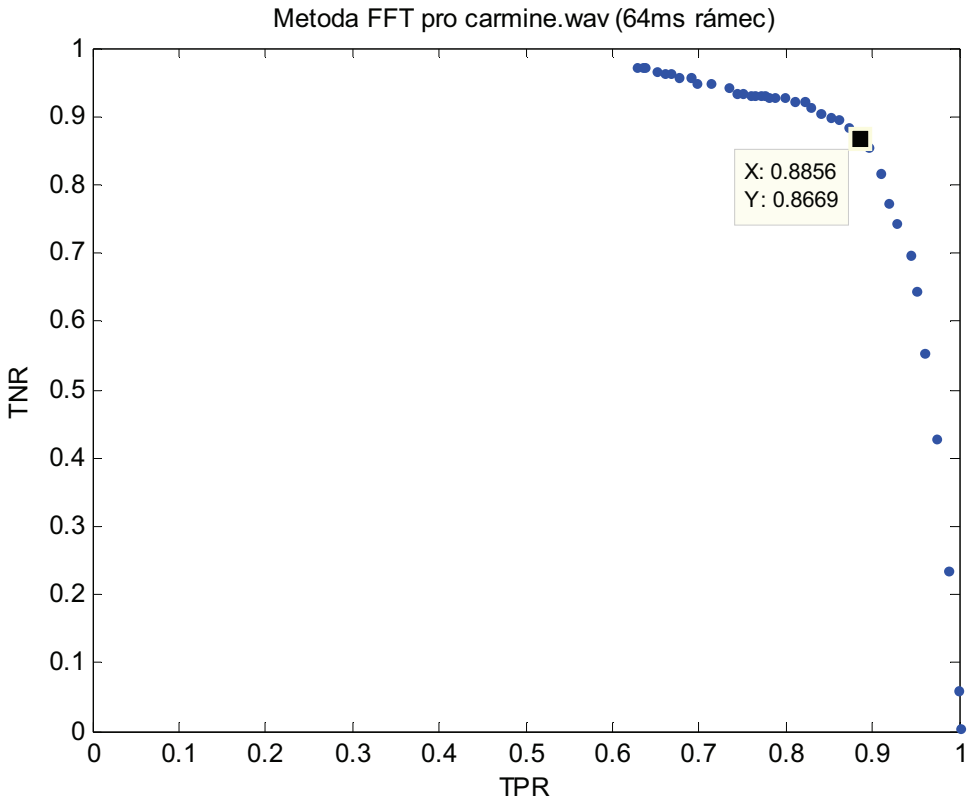

Metoda FFT pro paolo.wav (32ms rámec)
X: 0.7648
Y: 0.7718
TNR
TPR
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

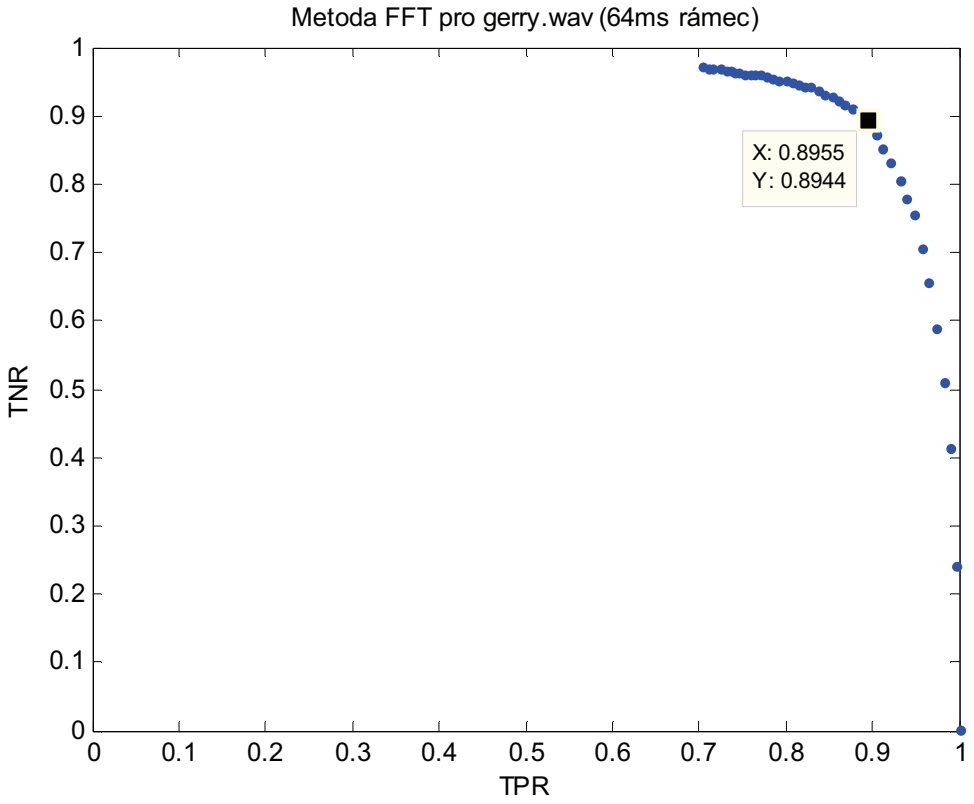

Metoda FFT pro tonyf.wav (32ms rámec)
TNR
TPR
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

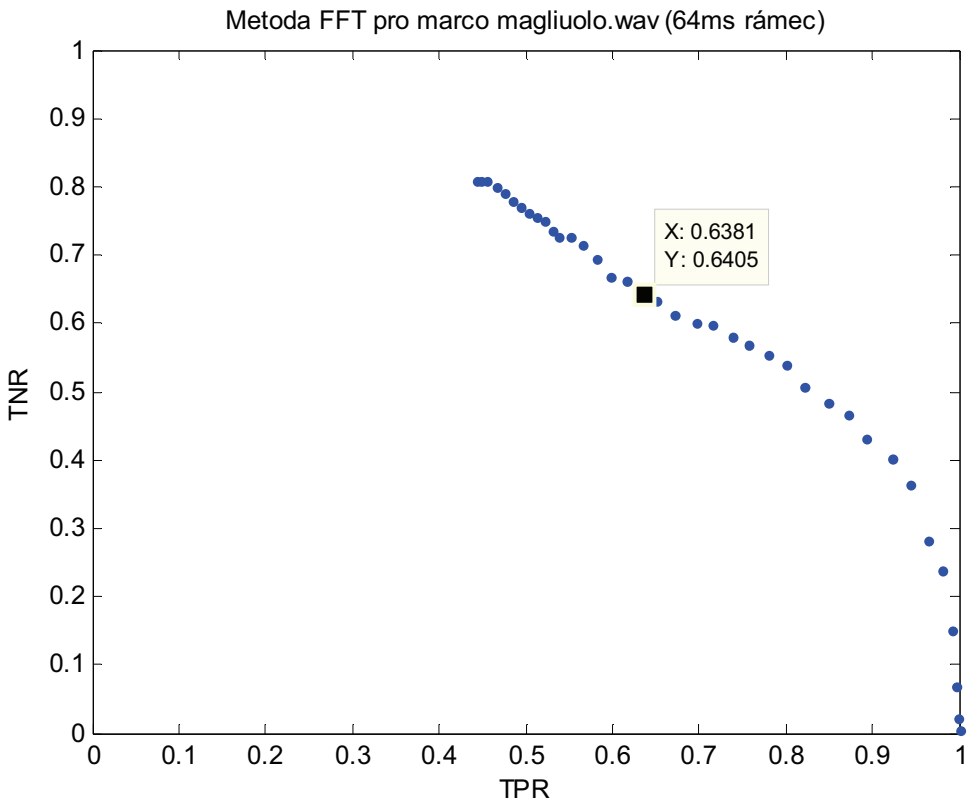
B ROC grafy metody FFT pro segment délky 64ms

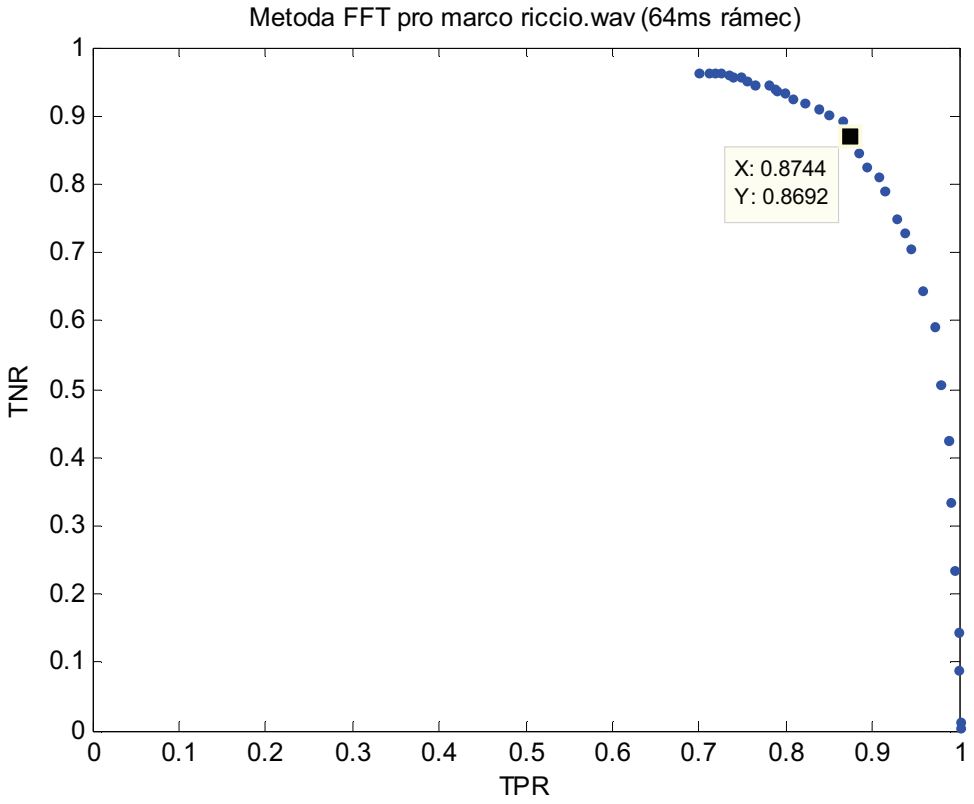


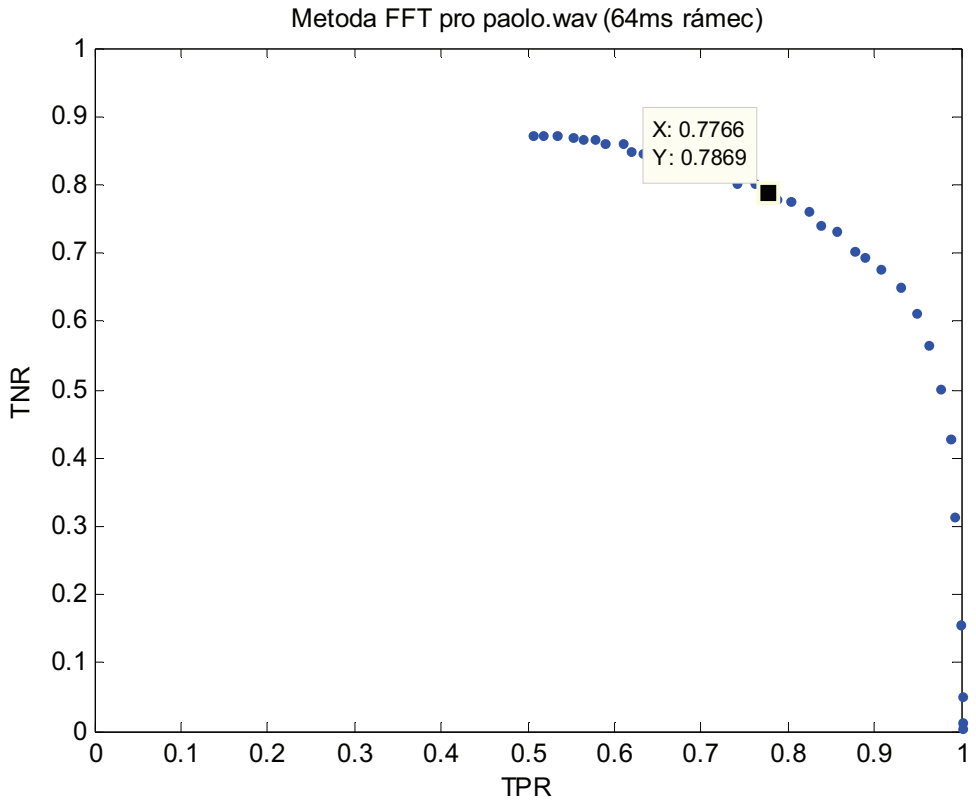


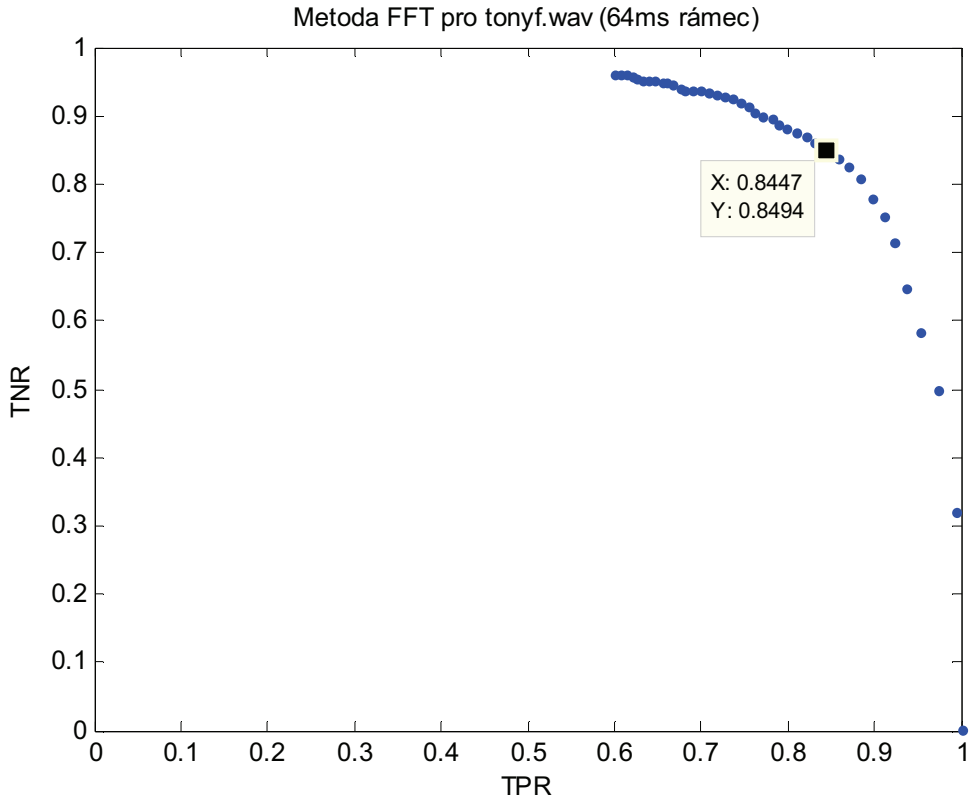

Metoda FFT pro carmine.wav (64ms rámec)
X: 0.8856
Y: 0.8669
TNR
TPR
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


Metoda FFT pro gerry.wav (64ms rámec)
X: 0.8955
Y: 0.8944
TNR
TPR
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


Metoda FFT pro marco magliuolo.wav (64ms rámec)
X: 0.6381
Y: 0.6405
TNR
TPR
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


Metoda FFT pro marco riccio.wav (64ms rámec)
X: 0.8744
Y: 0.8692
TNR
TPR
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


Metoda FFT pro paolo.wav (64ms rámec)
X: 0.7766
Y: 0.7869
TNR
TPR
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


Metoda FFT pro tonyf.wav (64ms rámec)
X: 0.8447
Y: 0.8494
TNR
TPR
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

C obsah přiloženého CD

Diplomova_prace.pdf	vytvořená diplomové práce
Nahravky	řečové nahrávky
Matlab	vytvořené programy
Segmentace	soubory s ruční segmentací